DEVELOPING HCI TECHNOLOGY TO AID IN COMMUNICATION RESEARCH FOR INDIVIDUALS WITH IMPAIRMENTS A LANGUAGE AND LINGUISTIC PERSPECTIVE

BY

JOSHUA HAILPERN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Associate Professor Karrie Karahalios, Chair, Director of Research
Professor Gary Dell
Assistant Professor Wai-Tat Fu
Professor Jiawei Han
Distinguished Professor Gregory Abowd, Georgia Institute of Technology

Abstract

This dissertation focuses on understanding the relationship between the design of computer interfaces and individuals with an impairment. We seek to explore how knowledge of an impairment can inform the design of systems for those with the impairment and those around them. As the design space of "individuals with impairments" is quite broad, we refine our focus to individuals with communication impairments (children with autism and adults with aphasia). Specifically, this work will examine how HCI and CS techniques can impact language development , empathy and the role communication impairment has on linguistic patterns and conversation quality. Each of these broad themes (language development, linguistic patterns, and empathy/understanding) are equally important aspects of communication. Within each of these themes, technology has a powerful role to play.

to my grandfather, Alfred Gunn
who taught me to take one $10 bill over five ones

Acknowledgments

Many people have contributed to my graduate education, my research and the work in this dissertation. I wish to express my gratitude to the members of my dissertation committee: Karrie Karahalios, Gregory Abowd, Gary Dell, Wai-Tat Fu, and Jiawei Han. I am grateful for the time they spent with me in countless consultations, feedback on publications, and reviewing this dissertation. I also wish to thank Karrie Karahalios, my advisor, for her long-term support of my research.

There are no words to express my gratitude to Gary Dell and Gregory Abowd. Their ideas, comments, suggestions, and advice have formed an integral part of who I am as a researcher and scientist. They have been a constant source of support and it has been a honor to work with both of them.

I would also like to thank the other collaborators who have aided me in my research over my graduate tenure. To this end, I would like to thank Brian Bailey, Aaron Benjamin, Laura DeThorne, Jim Halle, Julie Hengst and Jason Hong. I would also like to thank my collaborators and mentors in industry: Vicki Hanson, Nicholas Jitkoff, Tessa Lau, Loretta Guarino Reid, Shari Trewin, Fernanda Viégas and Andy Warr. For many of my projects, I have been privileged to work with other students. I would like to thank Brianna Birman, Abraham Fine, Joey Hagedorn, Andrew Harris, Reed LaBotz, Maurice Lam, Robert Sesek and Nikita Shkrob.

I would also like to thank those organizations that have help fund my research – The Department of Computer Science at the University of Illinois (Cultural Computing Fellowship), the University of Illinois (Dissertation Completion Fellowship), and Autism Speaks (The Dennis Weatherstone Predoctoral Fellowship). This funding has been critical for allowing me to complete this work.

I would like to also acknowledge Jon Ames, Jessie Cheng, Evan Coopersmith, Brett Daniel, Erik Hinterbichler, Craig Tashman, Aaron Weiner and all those people who provided emotional and moral support throughout my degree. Your friendship has been invaluable.

To truly thank my parents, I would need to write a book to list everything they have done for me. In many ways, my career is a surprising intersection of both of their backgrounds, and in this way, I have been lucky to always have two experts "in the family" that I can call for help (but not later than 9pm). Your strength has picked me up so many times when I thought I wouldn't make it through. And though I don't say it enough, thank you. I may not remember our trip to Disney World, but you **are** the best parents in the whole world-da. Thank you for teaching me to love to learn.

Marina, before I met you, I honestly felt alone. I was loved – I had great friends, amazing parents, but still by myself. But with you, I stop thinking of myself, and of "we." You challenge me, you push me, and you keep me level headed. As problems have come up, *we* solve them. And it is together that this degree and this document were finished. You have become to best part of me, just as the best part of this research we did together. Thank you Marina for being a partner, collaborator, and the best friend I could ever ask for. I can't wait to see what we do next.

My grandfather, Alfred Gunn, passed away in 2001, and eleven years later, I still think back very fondly on his memory. And perhaps I will always remember him with the eyes of a small child, but to me, he will always be a man who could do anything (including making M&M pancakes for breakfast). In so many ways, he remains larger than life. From him, I learned that nothing should get you down. He showed me that love and compassion are the stronger than anything. From him, I saw that nothing is ever broken, that things are just waiting to be fixed and cleaned up (with some rubbing alcohol). But perhaps most importantly, he taught me to not take things so seriously. With his rich laugh, everything became ok. Because of him, I have tried to always interject that humor (albeit dry) and laughter into every part of my life. So thank you for so many years. Thank you for placing such a high value on my education. And in so many ways, this document and my passion for life is because of you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Universal design focuses on designing our environment, products, and artifacts of our daily lives so that they are inherently accessible to both able-bodied individuals and those with impairments. A technology (in a broad sense) which is created to serve a specific deficit may often prove useful for the general population. Consider curb-cuts, which were originally created to allow individuals in wheelchairs to navigate sidewalks, but which also provide aid for able-bodied individuals with strollers, carts, or bicycles. Another example being the electric toothbrush that was first designed for patients with limited motor skills. Universal design provides insight into individuals with an impairment and the broader community. And by exploring and researching technology that impacts those with impairments, we can advance our knowledge of the human condition, develop new computer solutions, and create technology that can improve the lives of all individuals – with a large and broadly reaching impact.

This dissertation focuses on understanding the relationship between the design of computer interfaces and individuals with an impairment. We seek to explore how knowledge of an impairment can inform the design of systems for those with the impairment and those around them. As the design space of "individuals with impairments" is quite broad, we refine our focus to individuals with communication impairments (children with autism and adults with aphasia).

The specific goals of this work is to therefore illustrate human computer interaction solutions to help improve communication for individuals with impairments. Specifically, this work will examine how HCI and CS techniques can impact language development , empathy and the role communication impairment has on linguistic patterns and conversation quality. Language development is critical for communication. Empathy building is central for communication with individuals who have aphasia. And being able to understand the impact of communication impairment has on linguistic patterns and conversation quality is necessary for researchers to help improve communication. Each of these broad themes (language development, linguistic patterns, and empathy/understanding) are equally

important aspects of communication. Within each of these themes, technology has a powerful role to play.

The first part of this dissertation will focus on the individual with an impairment; specifically, how HCI and CS technology can help develop language and speech production in children with autism and speech-delays. As a child develops, acquisition of speech and language typically progresses with little or no explicit effort from parents, family, or doctors. Developmental disorders, such as Autism Spectrum Disorder (ASD), can significantly disrupt the natural development of social behaviors including spoken communication. One hallmark difficulty of children with Autism Spectrum Disorder (ASD) centers around communication and speech. Research into computer visualizations of voice has been shown to influence conversational patterns and allow users to reflect upon their speech. In this research, we explore the effects of audio and visual feedback on vocalization and speech in children with ASD. By presenting a child with a new interpretation of their vocalizations (though audio and visual feedback), we aim to provide them with an additional means of understanding and exploring their own voice. We have created two such solutions, targeting vocalization production, and shaping multi-syllabic speech. While this work has directly targeted the vocalizations and speech in children with ASD, tools have been developed with broader implications beyond the ASD community. These include: 1) creation of a coding guideline for quantitatively understanding how pre-verbal children interact with computers, and; 2) creation of video annotation tools for any research (or practice) based in video. Further applications of this research could impact children with speech delays and potentially even foreign language acquisition.

The second part of this document will focus on those around an individual with an impairment; in particular, discussing how HCI and CS solutions can increase empathy and understanding for individuals with aphasia. In addition, the second part of the document will examine how that same technology can uncover the impact of language impairments on linguistics and conversation quality for individuals with impairments and their conversation partners. This second part specifically targets individuals with aphasia, an acquired communication disorder, who constantly struggle against a world that does not understand them. This lack of empathy and understanding negatively impacts their quality of life. While aphasic individuals may appear to have lost cognitive functioning, their impairment relates to receptive and expressive language, not to thinking processes. We introduce a novel system and model, Aphasia Characteristics Emulation Software (ACES), enabling users (e.g., caregivers, speech therapists and family) to experience, firsthand, the communication-distorting effects of aphasia. By allowing neurologically typical individuals to Òwalk in anotherÕs shoes,Ó we aim to increase patience, awareness and understanding. Results from an evaluation of 64 participants

indicate that ACES provides a rich experience that strong increases understanding and empathy for aphasia. To build a system like ACES, we leverage a host of existing literature from the communication science and psychological literature. This research illustrates how the study of a disorder (specifically aphasia), can be used to help educate clinicians, family members, and friends about a disorder.

While these two parts of this dissertation are linked together by a common theme, their related literature and background have little overlap. To this end, The dissertation document itself is broken up into two distinct sections. Part I examines the role of HCI technology in helping to develop language and speech production skills in children with ASD and speech-delays. Part II examines the development of systems that can realistically (by passing a turing test) emulate the effects of a language impairment to build empathy, and increase understanding of the impact of language disorders on linguistics and conversation quality. Within each of these two parts are separate introductions, literature reviews and general summarization of the findings and implications. This dissertation then concludes with a final Chapter (in Part III) highlighting the broad findings and summaries of this work.

Part I

Autism, Language and Non-Verbal Data Collection

Chapter 2

Introduction to Autism, Language and Non-Verbal Communication Data Collection

As a child develops, acquisition of speech and language typically progresses with little or no explicit effort from parents, family, or doctors. Developmental disorders, such as Autism Spectrum Disorder (ASD), can significantly disrupt the natural development of social behaviors, such as spoken communication. Since language is "a unique characteristic of human behavior…[that] contributes in a major way to human thought and reasoning" (Lovaas, 1977), the communication deficits of children with ASD are likely to have detrimental effects on multiple aspects of their lives. The impact of this disability as well as its prevalence, estimated by the Center of Disease Control and Prevention (CDC) as 1 in 150 children (Center for Disease Control and Prevention, CDC, 2007), highlight the need for effective methods to facilitate the development of communication, including speech.

This first part of the dissertation focuses on the individual with an impairment; specifically, how HCI and CS technology can help develop language and speech production in children with autism and speech-delays.

The first set of chapters (4 - 8) present SIP, the Spoken Impact Project, which explores a new area of HCI: using real-time audio/visual feedback to facilitate speech-like vocalizations in low-functioning children with ASD. This work is grounded in HCI and behavioral science literature. We believe computer-generated feedback, generated from a child's vocalizations, can influence the vocalizations of children with ASD for communicative purposes by providing them with additional means of accessing information regarding parameters of their voice (e.g., pitch, loudness, duration). To facilitate the SIP research, I conducted 3 major pieces of research; Tools for Video Annotation (Chapter 4), A^3 Coding Guideline (Chapter 5), and the Experimental Testing of SIP's Visualization software (Chapters 7 and 8).

The remaining chapter VocSyl: Syllable Project (Chapter 9) focuses on taking SIP to one of the many possible "logical next steps," using visualization to teach specific langue productions skills (specifically

multi-syllabic speech). The work presented in this section is part of a larger research project that is currently underway, and will be continuing beyond this dissertation.

Beyond the experimental results, the contributions of this work are the creation of design guidelines to facilitate video annotation in an experimental setting based on the needs of researchers, a set of dependent variables to enable behavior analysis of non-verbal human-computer interaction, a demonstration of a new approach to ASD research (within the context of HCI research) and an initial understanding of how the SIP model could be further explored by the HCI community.

In order to motivate VocSyl and SIP, we first outline the general background, theory and existing state of both HCI and Communication literature (Chapter 3). When appropriate, we provide additional theoretical background within each sub section.

This work, together, highlights the great potential of voice visualization software for children with autism and speech delays; systems that sit at the intersection of human-computer interaction, special education, and speech and hearing science. At this intersection, our technological solutions can greatly impact quality of life (by increasing language acquisition) and quality of care (creating a more enjoyable, and theoretically a less stressful environment) while simultaneously developing new computer technologies that advance our understanding of how people interact with technology (and how to design new solutions for interaction).

2.1 Scope and Motivation of SIP

SIP explores a new area of HCI research focusing on the use of contingent audio and/or visual feedback to encourage sound production in low-functioning children with ASD. Without the development of techniques to encourage speech/vocalization, a diagnosis of ASD can have far reaching negative implications for a child's social, developmental and educational life.

Building on prior work, our focus on computer visulization in this population is unique. Most HCI visualizations research has focused on neurologically typical individuals (Schneiderman, 1996). ASD treatment research in HCI has targeted higher functioning children with ASD (Tartaro, 2006) but has failed to address the needs of non-verbal/low-functioning children with ASD. Though the literature in the behavioral sciences has explored this demographic, existing practices use low-tech alternatives such as PECS (Bondy and Frost, 2001), mirrors and echo chambers (Marshalla, 2001) or invasive procedures, such as electropalatography (Carter and Edwards, 2004). Our research begins with the basic question:

can real-time visual/audio feedback positively impact sound production in low-functioning children with ASD?

While there is discussion that high-functioning children with ASD should not be pressured to communicate vocally, this concern is not applicable to this vein of research. These children cannot communicate by any means (e.g., typing, signing or speaking). Teaching some form of communication is essential, though the method should vary according to individual preference and capabilities.

2.2 Scope and Motivation of VocSyl

VocSyl explores the role of computers in shaping multisyllabic speech in children with ASD and SPD. We developed a flexible and expandable software infrastructure. We included children with ASD and SPD who had limited speech-language skills in the design process. Additionally, we provide the research community a set of design guides for building future speech-language computer solutions.

Our main contributions are the creation of software, the documentation of the TCUID process, and the articulation of design guidelines for future software development that shapes multisyllabic production. Our research begins to bridge the gap between Hailpern's work (Hailpern et al., 2009b) on encouraging vocalization in children and Hoque's work (Hoque et al., 2009) on providing a game to practice full sentence production. It is important to note that teaching perfect speech or precise phonemes is outside the scope of this research – our software focuses on facilitating prosody and syllables. Further, developing or implementing 'the best' pitch or syllable detection algorithms is also outside of the scope of this work, and while our digital signal processing algorithms (see Table 9.1) may not always be 100% accurate, our implementations were consistent across users.

The TCUID process using children with ASD and SPD explores a new area of HCI research. It concentrates on the use of real-time visualizations of voice production to demonstrate and shape (by matching) multisyllabic productions. The focus of this research is, therefore, on the design and development of our software and the software design lessons-learned from our target population, rather than on quantitatively measuring the impact of our software (see Future Work for ongoing studies).

Chapter 3

Review of Autism Literature

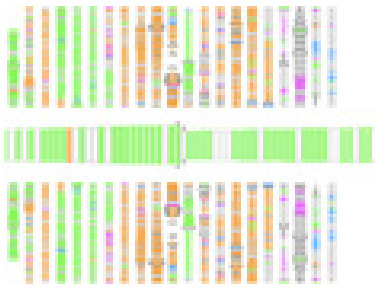
We present here an overview of existing research in Communication Treatment and HCI literature on Autistic Spectrum Disorder. Additional literature analysis is presented at the beginning of Chapter 4: Tools for Video Annotation, Chapter 5: A^3 Coding Guideline, and Chapter 9: VocSyl: Syllable Project, providing additional related work pertaining specifically to the topics covered in those chapters.

3.1 Computer Visualizations

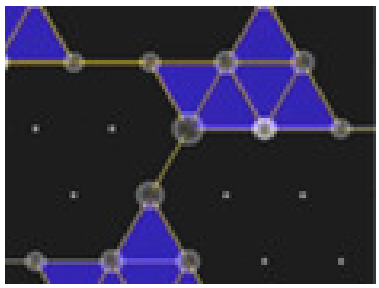
Computer based visualization systems have been providing users with new, and faster ways to understand large quantities of varied and complex data (Heer et al., 2007). See Figure 3.1 for examples of existing Visualizations. Work on Awareness Displays (Mankoff et al., 2003; Plaue et al., 2004) has shown the benefit of abstract representations, and their ability to provide continuous streams of data to users about their world, interests, and needs. Visualizations have also been used to impact teaching of complex concepts (Leung, 2006)by presenting visual and interactive representations.

Yet a major thrust of visualization research has focused on providing feedback to impact communication (Bergstrom and Karahalios, 2007b; Bergstrom and Karahalios, 2007c; Karahalios and Viegas, 1999; Karahalios and Dobson, 2005; Karahalios and Viegas, 2006; Levin and Lieberman, 2004; Viegas and Donath, 1999) and illustrate emotion (Ahn and Picard, 2006). These systems give the users a new understanding of their vocal interaction with others (from rate of speech to dominance in a conversation). These visualizations have the ability to alter people's behavior in real time. This research has shown that these visualizations directly impact the communication of said individuals. Thus, the potential impact of visualization systems is far reaching, for social, personal, and educational purposes. We hope these ideas generalize to other communities and those with communication impairments.

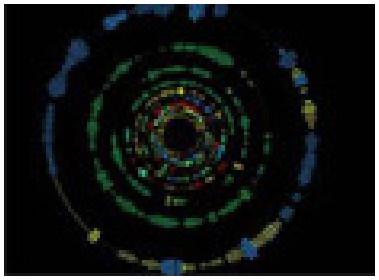
Some research, content, and text from this Chapter is reproduced from (Hailpern et al., 2009b; Hailpern, 2008)



(a) *Conversation Votes (Bergstrom and Karahalios, 2007b)*



(b) *Isochords (Bergstrom and Karahalios, 2007c)*



(c) *Conversation Clock (Bergstrom and Karahalios, 2007a)*

(d) *Telemurals (Karahalios and Donath, 2004)*

Figure 3.1: *Examples of existing visualizations*

3.2 Autistic Spectrum Disorder

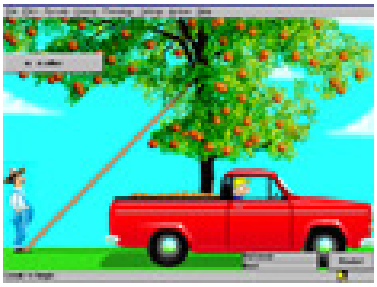
Kanner's 1943 description (Kanner, 1943) of 11 children with ASD documented this disorder in the scientific community. In the past 60 years, scientists and therapists have strived to better understand ASD and provide treatments to mitigate its many communicative and social difficulties. The ASD population is not a homogenous group. Many of the characteristic difficulties and developmental delays revolve around communication, empathy, social functioning, and expression. The Autism Society of America describes ASD as "insistence on sameness…Preference to being alone…spinning objects [and] obsessive attachments to objects"(Autism Society of America, ASA, 2007). While some children have limited impairment, those with a greater difficulty with social and communicative skills are considered low functioning.

3.3 Communication Treatment

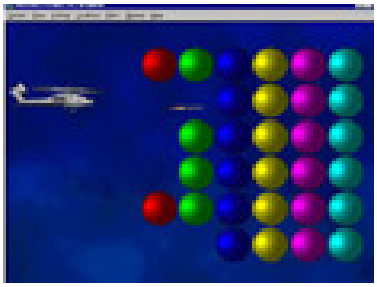
Since the 1960s, Ivar Lovaas' pioneering approach of "applied behavior analysis" has been used to help teach communication and social skills to children with ASD. The treatment focuses on extrinsic

(a) *PECS (Bondy and Frost, 2001)*

(b) *Go Talk (Attainment Company, 2008)*



(c) *Speech Viewer III (IBM, 1997)*



(d) *Visi-Pitch(KayPentax, 2008)*

Figure 3.2: *Examples of existing communication treatment solutions*

rewards (e.g., food or toys) for encouraging targeted behavior (Lovaas, 1977). Over time, rewards are slowly faded or removed resulting in more naturalistic behavior.

While the merits of this treatment have been documented for 30 years, this form of therapy has high financial and labor-intensive costs. Furthermore, frequent sessions requiring sustained attention and intense human-to-human contact can be anxiety producing (Baskett, 1996). This anxiety along with the detached/alone feeling of many children with ASD (Baskett, 1996; Kanner, 1943) causes difficulty for practitioners and subjects. Further challenges also concern generalization of these skills. Other forms of communication treatment (Greenspan and Wieder, 1997; Koegel et al., 1999; National Research Council, 2001; Woods and Wetherby, 2003) have been used to help develop social and communicative skills in children with ASD. Figure 3.2 illustrates four commonly used devices to assist in the development of communication skills.

3.4 HCI & ASD Research

Since the 1990s, the HCI community has examined how computers can aid in diagnosis of ASD (Hayes et al., 2004; Kientz et al., 2007; Kientz et al., 2006). In addition HCI has studies audio perception (Russo et al., 2008) and teaching human-to-human interaction to high-functioning children with ASD (Kerr et al., 2002; Lehman, 1998; Mohamed et al., 2006; Tartaro and Cassell, 2008). Elements of play have also been studied that demonstrate that technology/computers can reduce the apprehension caused by human-to-human interaction (Lehman, 1998; Michaud and Theberge-Turmel, 2002; Pares et al., 2005). Other HCI research (Bergstrom and Karahalios, 2007c; IBM, 1997)and technology-based behavioral science research (Bergstrom and Karahalios, 2007a; Barry, 1989; Shuster and Ruscello, 1992) outside of the ASD community has illustrated the use of computer solutions in the context of speech and communication therapy. Four of these solutions are illustrated in Figure 3.2 . These computer-based solutions tend to approach visualization through games, and controlling characters in environments. While these game-like solutions can be useful, they abstract away many of the vocal properties resulting in a lack of understanding what or how their voice may impact the visualization. Further, they require a cognition level that can understand a game, a goal, and discovering how to achieve that objective.

Speech recognition is a commonly used technique for computationally capturing speech for the purposes of archival and analysis. Due to the current limitations of speech recognition software (Nakagawa, 2002; Strik and Cucchiarini, 1999), the forms of speech detection are limited, especially for individuals with poor diction. Hence, technology must be designed to aid and supplement practitioners and researchers rather than replace them.

With this work, we explore methods and technology that can facilitate the speech and vocalization education process for children with communication skill deficits. Specifically we intend to use contingent visual and auditory feedback to (a) motivate and reward vocalization and (b) provide information about the acoustic properties of vocalizations. The following chapter details our approach for creating computer based feedback systems to encourage vocalization in non-verbal children with ASD.

Chapter 4

Tools for Video Annotation

Figure 4.1: *VCode and VData Suite of Applications for Video annotation*

Human behavior does not naturally lend itself to being quantifiable. Yet time and again, researchers in disciplines ranging from psychology to ethnography to computer science, are forced to analyze as if it was quantified. Those in human centered domains can now rely on video annotation to provide them with measures on which to draw conclusions. Unlike transcription, which is akin to what a court stenographer does, annotation is the marking of movements, sounds, and other such events (with or without additional metadata such as rankings). The emergence of technology as a tool to aid in video

Some research, content, and text from this Chapter is reproduced from (Hagedorn et al., 2008; Hailpern, 2008)

annotation has raised the possibility of increasing reliability, repeatability, and workflow optimizations (Burr, 2006).

Three notable limitations of existing video annotation tools are lack of support for the annotation workflow, poor representation of data on a timeline, and poor interaction techniques with video, data, and annotations. This chapter details a set of requirements to guide the design of video annotation tools. Our model is the direct result of an analysis of existing tools, current practices by researchers, and workflow difficulties experienced by real-world video coders. By understanding what data researchers are looking to gather, and the shortcomings of existing techniques and technology utilized by coders, we believe that we have created a framework for video annotation that can reach across disciplines. Our model is demonstrated through the design and construction of our new system VCode and VData (Figure 4.1); two fully functional, open-source tools which bridge the video annotation workflow.

The primary contribution of this chapter is the set of design requirements for facilitating a system conducive to video annotation. Specifically, we demonstrate how a system could be designed and built to meet these requirements through a set of carefully designed interfaces and graphical representations of data.

4.1 Related Video Annotation Work

4.1.1 Video Coding in Practice

The analysis of human behavior is a study that dates back hundreds of years. This has ranged from anthropologic ethnography (Clifford et al., 1986) to psychological evaluation. As technology has developed, the use of video and creation of annotation techniques have aided researchers by providing a referable document that can be used as evidence to back up claims and observations made (Lee, 1974; Retherford et al., 1993; Rosenblum et al., 2004). These techniques involve detailed event logging on paper, specifying feature such as durations, ratings/levels, and time-stamps (Leadholm et al., 1992) . To ensure a reliable set of data from annotation, researchers perform agreement calculations between coders (Berry and Mielke Jr, 1988). This agreement is utilized throughout the data gathering process (by testing some small percentage of data segments to ensure consistency throughout), but also during training of coders (to decide when they fully understand what events they are looking for). There are many techniques for calculating agreement including Cohen's Kappa (Conger, 1985; Kazdin, 1982), Cochran"s Q-test, and Point-By-Point Agreement. Regardless, the management of data with traditional means is considered "cumbersome" (Retherford et al., 1993).

4.1.2 Video Coding Tools

Digital annotation tools have demonstrated significant benefits from simple copy/paste and undo to increased quality of coding by the facilitation of multiple passes on video and graphical representations (Burr, 2006; Quek et al., 2003). A timeline is commonly utilized in these tools and is familiar without extensive training (Costa et al., 2002). Existing research also indicates that presenting coders with secondary or sensor data on a timeline helps them outperform coders without sensor data (Burr, 2006). Increased accuracy, quality, and speed not only enhance the data collected, but also allow for more annotation to be conducted in the same amount of time. In addition to the computational benefits of digital annotation tools, they also provide a controllable mechanism for deferent forms of reliable video playback. (Quek et al., 2003).

One critical limitation of existing tools is poor representation of data on a timeline and utilization of screen real estate. For example, the VACA solution, while utilizing minimal screen real-estate by condensing all annotations to one large easy to read track, presents a problem with overlapping and simultaneous events (Burr, 2006). The VisSTA solution takes the contrary approach by showing many small vertically tiled tracks. Though this allows for a good comparative view, reading individual annotations & holistic interpretation is difficult due to scrolling (Quek et al., 2003). These and other existing solutions have not successfully dealt with this problem (Abowd et al., 2003; Costa et al., 2002; Johnson, 2007; Kipp, 2007; Noldus, 2007; Ramos and Balakrishnan, 2003; SaySoft, 2007; Strik and Cucchiarini, 1999).

Another limitation of current annotation tools is poor interaction techniques with video and data. Though robust functionality is provided for playback, controls can be cumbersome & overly complex, (e.g. (Quek et al., 2003)). Too many windows resulting in an over-saturation of information, imprecise video interaction & annotation or rigid, inaccessible marking interfaces (e.g. (Kipp, 2007; Ponceleon and Srinivasan, 2001; Quek et al., 2003; Ramos and Balakrishnan, 2003; Strik and Cucchiarini, 1999)). Each of these are common stumbling blocks which could result in unreliable data.

One last limitation is lack of support for the full annotation workflow that follows researcher from experimentation to data analysis. The larger the degree of support, the smaller the chance of error, and the more efficient the data gathering, collection, and analysis process becomes. This workflow was created through discussions with coders, researchers and an examination of the existing literature. The following are the 6 steps of the video annotation workflow:

1. collect video
2. create segments to code

3. train coders/demonstrate reliability
4. gather data
5. perform regular checks on reliability & discuss discrepancies
6. perform data analysis.

Many tools support small portions of this workflow (i.e. simply facilitating segmentation, annotation, or reliability (Burr, 2006; SaySoft, 2007; Studiocode Business Group, 2007)), but with each break in the process researchers can become delayed. Without export/import data reentry is required. Technology is situated to optimize this process. Researchers have also explored dialogue transcription (Johnson, 2007; Salt Software LLC, 2007; Studiocode Business Group, 2007), tagging (Abowd et al., 2003; Studiocode Business Group, 2007), scene based automatic annotation (Chen et al., 2002; Ponceleon and Srinivasan, 2001), automatic event logging (Banerjee et al., 2004), and object of focus identification (Bertini et al., 2004). This chapter contrasts these other foci by demonstrating techniques for supporting human based annotation of events that occur in video.

4.2 Interviews and Collaboration

To gain a deeper understanding of methods, analysis processes, bottlenecks, and types of data needed for effective video annotation software, we maintained an active dialogue with researchers (in Special Education, Speech and Hearing Sciences, and Computer Science) who use video annotation, conducted informal 40 minute interviews with two experienced video coders, and refined functionality through dialog with current users of VCode and VData. Existing tools for video annotation may address a subset of the below described requirements, however, our system more fully satisfies all of them.

R1: Facilitate Coding Workflow: The coding workflow consists of; (1) establishing video clips and coding guidelines, (2) intense training of coders and checks for reliability, (3) annotation of videos, (4) weekly reliability checks on annotated videos, (5) repeat 3 and 4 ad infinitum, (6) analyze data in statistical packages. Tools targeting video annotation should attempt to optimize the transition between steps in this workflow.

R2: Video, Annotations, and Coding Guidelines should be presented in a synchronized manner: Interviewees described their coding process centering around analog video on a TV-VCR device, annotating in a Microsoft Excel file, and referencing lengthy code guidelines. Due to the visual separation between annotations, source material, and video, coders had great difficulty during reviews.

R3: Capture Appropriate Data: Researchers and existing literature indicate that there are different types of data that are collected through the annotation process: counting events/occurrences, determining duration of events, assigning levels, values, or ranking to events, performing phonetic transcription, and general commenting (Kipp, 2007). Effective interfaces must provide methods for capturing these conceptually different data types while preserving each of their unique nuances.

R4: Additional data should be displayed to coders: Effective annotation tools should allow researchers to provide additional data to coders to aid in their assessment of video; for example, a volume histogram of the current video, sensor/log data collected in tandem to the video capture, or annotations made automatically or from another source. Displaying additional data points has shown to increase the accuracy of coded events (Burr, 2006). Further, annotation software should facilitate the management of multiple video streams to get the most accurate ”view” on the session, and thus produce the most accurate data (Quek et al., 2003).

R5: Allow multiple forms of playback: Researchers mentioned that continuous playback is not always the preferred method of analyzing a video. Often multiple modes of playback are utilized; continuous or standard playback, continuous interval playback (play for N seconds, then stop), and skip interval playback (jump N seconds, then stop). This allows the video to be divided in to smaller segments for annotation of events that are more difficult to pinpoint (i.e. when a smile starts or ends) (Quek et al., 2003). Though conceptually simple, manipulations of video using a standard VCR was described as ”annoying” and ”a mess” due to hand eye coordination and repeatability issues.

R6: Agreement calculations should be easy and manipulatable: Regardless of agreement technique used, researcher expressed a frustration in attempts to calculate inter-observer reliability. Specifically, existing solutions were limited to importing data into a statistical software package for calculation or calculating them by hand. Video annotation tools should provide quick & easy reliability calculations for individual variables, as well as overall.

R7: Provide functionality for visual, graphical and contextual review of annotations: In interviews, coders lamented the process of ensuring reliability on a weekly basis; as it consisted of searching through printouts of a spreadsheet for discrepancies. Specifically, by lacking context in this spreadsheet coders found it difficult to recognize what a given coding mark referred to due to the lack of synchronization with video. By providing a visual, graphical way to review annotations

Figure 4.2: *Two Coders and Researcher reviewing a coding session.*

(in the context of the video) coders would be better able to justify the decisions, determine the correct solution, and save time identifying the errors.

4.3 VCode and VData

VCode and VData are a suite of applications which create a set of effective interfaces for the coding workflow following the above design requirements. Our system has three main components: VCode (annotation), VCode Admin Window (configuration) and VData (examination of data, coder agree ment and training). The interaction with VCode and VData is demonstrated in Figure 4.2 - Figure 4.8 in which two coders are marking a video of a child in an experiment, and checking the agreement between their annotations. The reader should note our solution is only one possible implementation of the design requirements, and that these requirements could be applied to improving existing video annotation software.

4.3.1 VCode

The VCode application (Figure 4.3) is designed to provide researchers with an effective way to obtain reliable data from an observational research video. By allowing researchers to present multiple video

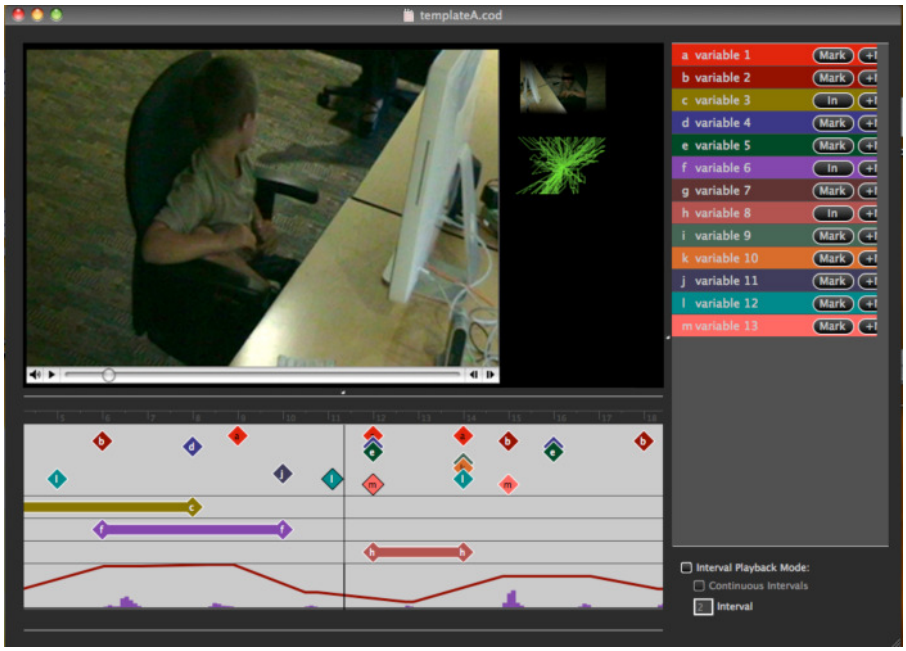


Figure 4.3: *VCode Interface*
The VCode application graphically represents the behaviors being coded as marks on a timeline. It is easy to see correlation between the marks on the timeline and sensor data displayed below.

streams in addition to other sensor data (e.g. log data, annotations from other software, or signals recorded by a computer/monitoring device) the coder can make the best annotation decision possible.

Video: To facilitate multiple video streams VCode presents one main video at full size, and a dock with other streams playing in real time. When a docked stream is clicked on, it repositions itself into the main video window, while the video which was the previous focus, scales down to the dock, thus equating visual importance with relative size and visual weight.

Variable List: To provide coders with a persistent list/understanding of their task and objective, VCode presents them with a list of each variable they need to annotate. This list (presented along the right hand side of the VCode window) displays each variable"s name, a unique color associated with that variable, a keyboard hot key for placing marks, as well as UI buttons to facilitate mark placement.

Events: When annotating a video, two different classes of coding events emerge: ranged and momentary. A ranged event is one that extends over a period of time (marking action start and duration). Momentary marks have no duration, and thus represent one specific moment in time. Each mark appears with it"s associated color and it"s keyboard hot-key letter (see Figure 4.5). These color and hot key based identifiers make it easy for coders to quickly assess which events represent which variable.

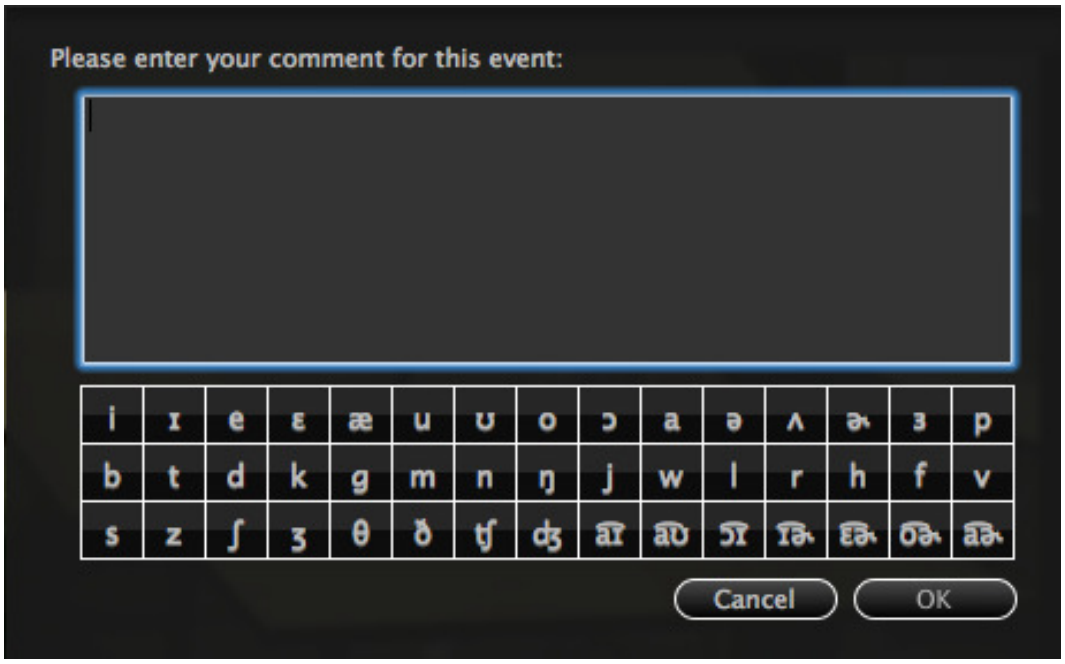


Figure 4.4: *Comment and Transcription Interface*

Comments can be attached to any mark (See Figure 4.4), allowing additional observations, levels/ranking, or phonetic transcription (through onscreen phonetic keyboard). Any mark with a comment has a inverted outlines to signify that it has a comment attached. Figure 4.3 and Figure 4.5 show a ranged event representing the length of time which a child is making a sound, with additional momentary marks at the start noting other features of the child"s state of being).

Timeline: The timeline is the heart of VCode. It is modeled after the moving timeline one might find in a video editing application (e.g. iMovie, Final Cut Pro, etc.). Events, graphically represented by diamonds, appear in a spatial linear fashion to sync with the video. Once an event has been placed on the timeline, dragging, clicking, and double-clicking can graphically manipulate the mark. The standard solution for dealing with large numbers of tracks or variables is to provide a vertical scroll bar or overlay tracks. Rather than limiting the amount of information on screen by scrolling, tracks representing momentary events are "stacked," such that they vertically overlap. This optimizes usage of the screen while still providing enough area for track isolation and selection, even under dense data conditions.

Ranged event tracks are unable to benefit from this stacking optimization because of the more complicated interaction for manipulation and thus are vertically tiled. Researchers can present video volume,

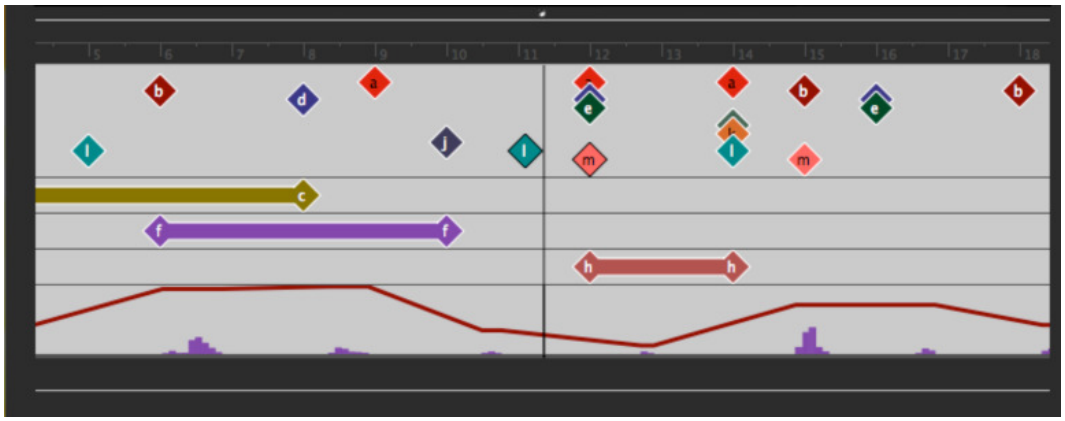


Figure 4.5: *VCode Timeline*
Each event is marked as a diamond, with the corresponding hot-key being represented on the mark for easy identification. A time stamp runs along the top of the timeline, with a play head represented vertically.

sensor data, software log data (from Eclipse or Photoshop for example), and even other annotations to the coders. This additional information is presented graphically to the users by bar, line, or scatter plot. This secondary data can allow coders to annotate data captured by other sources than the video streams, as well as provide additional context to their code. For example, should a coder be instructed to mark when a certain noise occurs, he can line the mark up with an audio peek, rather than estimate it and be concerned with reaction time.

Interaction: Annotations can be inserted into the timeline via UI buttons or keyboard hot keys. In order to increase association between hot-key and timeline, timeline marks have the letter of their hot-key on them. To optimize the typically complex transport controls we isolated the key activities that coders need execute and provided controls limited to play/pause buttons, coarse and fine grained playhead positioning, and step controls. The three modes of playback outlined in R5 are available.

4.3.2 VCode Administration Window

To ensure consistent configuration between coders and sessions, all administrative features are consolidated in a single window. The expected workflow is such that a researcher would setup a single coding document with all the variables to be used on all the videos. This template would then be duplicated (with media and log files inserted for each trial). The main task the Administration Window (Figure 4.6) is to facilitate is the creation of tracks, used to code data. Researchers can add, remove, and reorder tracks that appear in a list format. The name, color and hot key of each tack can be set through this

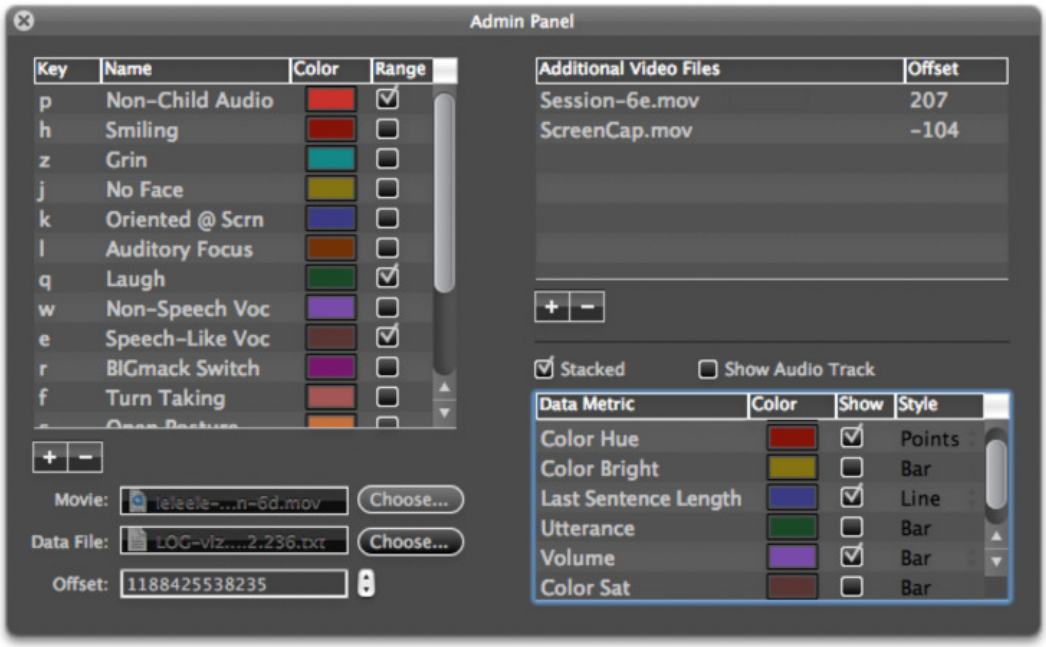


Figure 4.6: *VCode Administration Window*
The code is specified in the Administration Window along with the different video angles, screen capture, and log data.

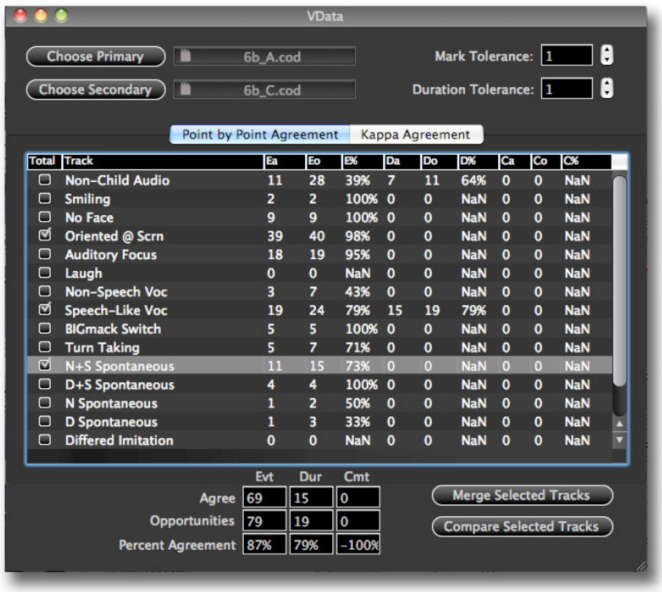


Figure 4.7: *VData Interface*
Later, analysis is performed on independent codings of the same video using VData

list presentation. Tracks can be enabled as ranged events through a check box in this interface. The Administration Window is also where a researcher specifies videos and data file to be coded, as well as secondary data for contextual annotation. These elements are specified and synchronized through a drag and drop interface, all of which is hidden from the coder to prevent configuration corruption.

4.3.3 VData

Critical aspects of the video coding workflow (training, reliability, and accuracy) revolve around demonstrating agreement between coders. VData (Figure 4.7) is a separate executable application specifically targeted to aid researchers in training and agreement analysis of coded data produced in VCode.

Multi Coder Analysis: By loading two VCode files into VData, tracks are automatically loaded into the main data table that presents all the raw data (opportunities, agreements) and percentage agreement for point-by-point agreements calculation. For each event (momentary or ranged) an opportunity is said to occur when the primary coder makes a mark. If the secondary coder also makes a mark within a specified short interval, the marks are said to agree. A percentage is calculated

from agreements/opportunities for easy interpretation. A tolerance variable is also present to (1) accommodate for variability in the mark placement by the coders, and (2) recognition that there is no quantization of marks beyond the granularity of the millisecond timescale, a property of the system. VData also provides agreement for ranged events and annotations in a similar fashion.

In addition to point-by-point agreement calculations, VData also supports Kappa Calculations by switching tabs. Much like point-by-point agreement, all of the raw data needed to perform a Kappa calculation is presented in the application. Kappa calculations is based upon the notion that within a given time span (the video duration), there are a discrete segments called opportunities. By opportunity we mean, that in a certain period, an observation can be made (aka something DID or DID NOT occur). When calculated, Kappa takes into account the probability that two coders made a mark at the same location by chance. The presence of a annotation in VCode, means an event occurred. The absence of a mark means it did NOT occur. Kappa examines the number of agreements between two coders for events occurring and not occurring. As a result, this form of agreement calculation only applies to those tracks that were annotated using interval playback mode.

It is not uncommon for multiple tracks or variables to by measuring slight variations on a theme (e.g. smiling vs. large smile vs. grin), thus VData implements a track-merging feature which allows opportunities on two distinct tracks to be treated indistinguishably. The resulting hybrid track can be used to see all the same raw data in addition to the percentage agreement. This new track can be treated as any of the original tracks from the VCode files.

For a holistic view, researchers can select tracks to be added into a total agreement calculation. In other words, if analysis determines that a single track is not reliable or it is determined that a given track will not be used in the future, it can be easily excluded from the total agreement calculation.

Conflict Resolution & Exporting: We have optimized coder training and reliability analysis by providing a graphical mechanism to directly compare annotations of two coders. VData can create a VCode session containing specific tracks of two individual coders for side-by-side comparison (Figure 4.8). Both coder"s marks appear side-by-side along side the source video, both with similar colors (e.g. light green, and dark green). The visual, side-by-side, representation of the data makes it easy to recognize systematic errors in context and detect differences between two coders markings. This reduces the time necessary to locate discrepancies and discuss the reasons why they might have occurred. It is necessary to keep records of these agreement analyses performed with VData by text export. Maintaining export at each stage of the process provides additional transparency and maintains traceability of results that come out of the system.

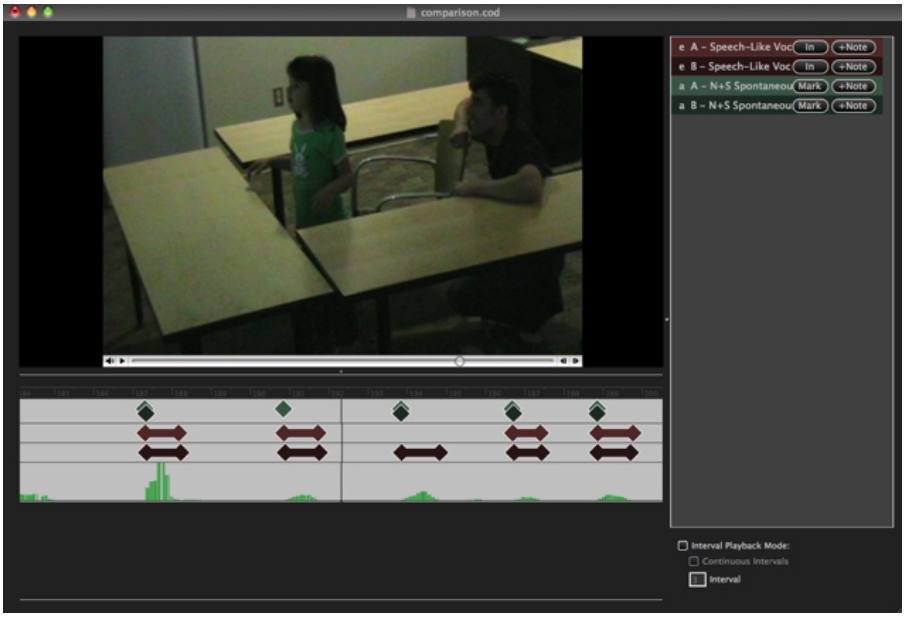


Figure 4.8: *Reviewing Annotations in VCode*
A track with low agreement can be reconciled by viewing the results of two coders side-by-side in VCode, thanks to the capabilities of the VData analysis tool.

4.3.4 Implementation

VCode and VData were implemented in Objective-C using the Cocoa Framework for Mac OS X 10.5. VCode supports all video formats and codecs supported by QuickTime to enable wide compatibility with available video files.

4.4 Meeting the Requirements

To ensure Video, Annotations, and Coding Guidelines are presented in a synchronized manner, VCode provides a unified interface containing the target video, a timeline with graphically represented annotations (ranged event, momentary event, or comment depending on data metaphor) (R3), additional tracks of signal data (to increase accuracy) (R4), and a list of coding guidelines (R2) which place marks and stand as a visual reminder. Three forms of video playback (continuous/standard, interval playback, skip interval playback) are available via check boxes on the main VCode window to allow easily switching between modes of playback (R5). VData provides a dynamic interface for real time calculation of multiple agreement values to facilitate easy and dynamic agreement calculations (R6).

Through the transparent calculation process, researchers can see both the raw data, and the percentages side by side for easy judgments about the reliability of data collected. Upon request a visual, graphical and contextual review of annotations for both agreement review and training is supported (R7).

Finally, the Coding Workflow (R1) is encouraged through VCode"s template model in conjunction with the separate VCode Administration Window for easy set up and configuration. Training, data collection, and inter-coder agreement are enabled through a tight collaboration between annotation environment and agreement analysis. By consistently providing data export, researchers can be assured that any information annotated by coders can be easily extracted and exported into the statistical analysis tool of their choice.

4.4.1 Initial Reaction

To evaluate our system in a cursory fashion, we conducted an informal series of interviews with several coders that used our system during the course of an independent study. Analysis using VData showed inter-observer agreement was good and provided valuable coded data for the study. In general, comments from the coders were positive, especially when comparing the VCode system to non-computerized methods. One coder wrote: "The software was easy to use in general, and cut down on coding time." Several features of VCode stood out in their comments; color coding of tracks provided direct linkage between events on the timeline and the description panel, the correlation between files was clear to see during review, sensor data helped anticipate events and accurately code them. It was also noted that the sensor data provided reassurance that what they had noticed in the video was actually correct.

In addition to these positive marks we uncovered several shortcomings of the interface. The seemingly low-resolution bar-graph of volume data left coders unsure where precisely to make their mark. Because the elements of this graph are relatively wide, it appears especially coarse in comparison with the precision with which one may place a mark on the timeline. A spectrogram was suggested as an alternate visualization of the audio data that could help understand sound and video.

From a quantitative standpoint, the time required to annotate video for SIP drastically improved over the course of the experiment. At the beginning of their coding experience, coders took roughly 40 minutes per 1 minute of video footage (to annotate all variables listed in Chapter 5: A^3 Coding Guideline). By the end of the coding period, coder"s time was reduced to about 20 min per 1 minute of footage. Coders attributed this improvement to the ease of use of the annotation system.

Overall, results from these interviews and raw speed improvements, are very encouraging and suggest a more formal study to determine if performance improves in the same way that coders stated that they felt as the tool lowered the amount of time necessary for coding.

4.5 Other Applications

The VCode framework has additional applications outside of the HCI and Behavioral Science research communities. In many clinical settings (e.g. speech pathology), practitioners record their sessions with a subject, and post-hock analysis of the video to assess subject"s progress. Many of the same features provided in VCode can facilitate analysis of clinical sessions, providing therapists with a quick, reviewable and accurate tool for documenting progress and behavior. Further, the ability to export data allows clinicians to graph progress made, and be able to show the client (or guardian) the beneficial effects of treatment.

4.6 Summary and Future Improvements

Video annotation tools can be valuable to researchers by enhancing the annotation process through increased reliability, repeatability, and workflow optimizations. However, many existing solutions do not fully address all the needs of researchers and coders; effective representation of data on a timeline, efficient and robust interaction techniques with video and data, and support for the full video annotation workflow. Our research has provided many contributions in addressing these weak points.

We create a set of design requirements based on existing literature and annotation techniques, interviews with experienced coders, and discussions with researchers in multiple disciplines. Based on these investigations, we implemented a system, VCode and VData, that largely satisfies the requirements we outlined. These systems were then used in SIP, and coders were interviewed concurrent with and after using the software, and their reactions were solicited. Our model demonstrates how video annotation software, for many disciplines, can be enhanced to meet the needs of both researchers and coders.

As of the date of this thesis, 15,111 downloads of the VCode software package from multiple universities and countries around the world. Of note, the VCode package has been used at Departments of Computer Science, Human Computer Interaction, Psychology, Cognitive Science, Neuroscience, Education, Information Systems Engineering, Design Research, and Library & Information Science.

From the reaction of the coders, as well as our own assessment of VCode and VData, we have many directions of possible future work. One avenue is creating a database or networked system in order to facilitate remote access to content, and management of coding objects and assignments for individual coders. It is foreseeable that the system could be extended to a tool to prepare coding files; assist in dividing up raw footage, syncing data to video en masse, and other automation hooks. This could leverage some of the other existing work in automatic video segmentation. Lastly, we hope to address some of the concerns of our coders, including creating a richer set of data visualizations.

With such a tool sets available to researchers, more complex sets of variables are available to research to code. The follow chapter details the development of a new coding system that leverages many of the features of VCode and VData. This new coding guide allows researchers to assess the impact of computer based feedback systems, on nonverbal subjects through behavioral analysis.

Chapter 5

A^3 Coding Guideline

Work conducted in HCI to date has explored diagnosis (Kientz et al., 2007), play (Michaud and Theberge-Turmel, 2002; Pares et al., 2005), audio perception (Russo et al., 2008), and interpersonal skills for high functioning children with ASD (Tartaro, 2006). Although this work is greatly beneficial, the potential of technology to facilitate vocal development in lower-functioning children with ASD has received little attention. As a result, there is little work in the HCI domain that provides a model for how to quantitatively assess the impact of an intervention to encourage speech with low-functioning children with ASD using HCI. It is essential, that whenever assessing a novel design approach, to have tools and methodologies (in the case of this research, in particular, and for assistive technologies, in general) to document that design and those techniques, so that the value of the novel approach can be evaluated and compared to the state of the art.

We propose A^3 (pronounced A-Cubed) or Annotation for ASD Analysis to quantitatively assess a set of dependent variables identified through the digital video annotation process. Through the application of A^3 in this research context, we demonstrate the inter-rater reliability of the annotations, as well as directions for its improvement. Because we are required to rely entirely on subject behavior, rather than on feedback provided by subjects (due to the nature of ASD), the creation of such an assessment tool as A^3 is critical for evaluation of technology used by the ASD community. The contribution of this chapter is in the demonstration of a new coding system grounded in theory from multiple domains and demonstration of its reliability when applied in the context of SIP, an experimental study.

When working with non-verbal subjects (those who do not use language to communicate) interacting with a computer-based intervention system, one wonders how to assess their behavior in relation to the computer system (e.g. "how does a computer system increase speech-like vocalizations or turn taking?" or "are subjects engaged with the visual and/or auditory aspects of a computer system?"). Unlike traditional HCI experiments, this subject pool cannot engage in talk/think-aloud protocol,

Some research, content, and text from this Chapter is reproduced from (Hailpern et al., 2008; Hailpern et al., 2009c; Hailpern, 2008)

answer a questionnaire, or have a meaningful discussion on the merits of the technology. Traditional psychometric assessments and standard HCI task-based studies do not easily apply to this group of subjects. As a result, researchers are forced to rely upon analysis of the subjects' behavior to assess their interaction, engagement, and reaction to computer-based interventions. Since no formal parent report or normative tools are currently available to directly assess a child's vocal response during computer interaction, we focused our methods on direct observation. However, this raises additional questions such as what behaviors should be considered, how to sample, and thereby, how to assess behavior.

This chapter illustrates how theory from multiple disciplines can be brought together to create a set of dependent variables for use in video annotation, which can be used to assess the interaction between non-verbal subjects and computer-based interventions. This coding guideline called A^3 (pronounced A-Cubed), or Annotation for ASD Analysis is the primary contribution of this research. A^3 examines the interactions of non-verbal subjects with computer feedback in terms of engagement and attention, specifically focusing on speech-like vocal behavior. This chapter presents the full coding guideline and justification grounded in existing literature. By basing each variable on pre-existing work, we leverage the large body of literature that already establishes the usefulness and application of each variable. Further, A^3 is demonstrated in an experimental context through agreement analysis of over 1200 minutes of experimental video footage.

In addition to the creation of A^3 and reliability demonstration in an experimental context, we examine the existing body of literature and the current state of computer vision and voice detection. We apply state-of-the-art capabilities of computer technology to the A^3 guideline/annotation process and detail how, and to what degree, the process may be automated. This analysis is grounded in the technology capabilities of today, rather than what computer automation may be capable of in the future. These solutions may be integrated into the annotation process, reducing the burden on coders and researchers. Because A^3 is based on diverse literature from a wide span of disciplines, the applications to non-verbal subjects interacting with computer-based systems are far reaching. This chapter concludes by discussing how, and where, the A^3 guideline may be used.

Though this work is situated in an experimental context studying Autistic Spectrum Disorder (ASD), the application reaches beyond this population. A^3 can be utilized in diverse methodologies (e.g., single subject and group design, short term and longitudinal studies) examining a wide range of populations including individuals with limited speech output (e.g., verbal apraxia and selective mutism) and across multiple domains (e.g., HCI, behavioral science, and clinical practice). As researchers, it is incumbent upon us not only to construct tools, but also to reliably test and measure the performance of our solutions. It should be noted that A^3 was not designed to, or has been tested to, be used as a tool for

diagnosing autism, or performing general behavioral assessment. A^3 provides an operationalized system, grounded in existing literature from multiple disciplines, for quantitatively assessing the behavior of non-verbal subjects interacting with computer-based interventions. Investigators and clinicians are encouraged to adopt its use to their own purposes as guided by their own knowledge of best practice. Because A^3 focuses on vocalization and interaction of non-verbal participants with computers, we believe that the coding guidelines outlined in this chapter can be applied to many populations, experimental designs, and contexts.

5.1 Experimental Context

This chapter is situated, and demonstrated, in the SIP experiment that lasted over a 12-month period described in Chapter 6: Spoken Impact Project (SIP), which focused on children with Autistic Spectrum Disorder (ASD). ASD is a developmental disorder exemplified by delays in empathy, basic social interaction, communication, and language use/acquisition. As the name connotes, it is not a monotonic diagnosis, but rather a spectrum with ranges in severity and symptom presentation. The focus of this research was to examine the effect of computer-generated feedback on the behaviors of five non-verbal children with ASD. The feedback generated (audio and visual) was directly based on the individual participant's vocalizations. The interaction of each participant was assessed in terms of his or her engagement with the computer/feedback, attention, and vocal behavior.

The five participants were exposed to a variety of computer-based feedback systems over six sessions that were spaced approximately one week apart. Sessions lasted about 40 minutes, and consisted of eight trials, each about two minutes in length. With parental consent, sessions were video taped, totaling over 1,200 minutes of video footage for data collection. The experimental context focused on non-verbal low-functioning individuals with ASD who varied in age (3-8 years), but could all be characterized at the prelinguistic stage in terms of intentional communication.

5.2 Existing Literature

At the outset of the original experiment, researchers reviewed existing literature across multiple disciplines in HCI and Behavioral Sciences. Researchers concluded that no one source provided a set of dependent variables for observation of engagement and vocalization of non-verbal subjects in computer-based intervention contexts. Although the literature has established methods for coding the

behaviors of non-verbal subjects, no system has been designed for explicitly examining the interactions between non-verbal subjects and computer-based intervention. We present a brief summary of existing literature pertaining to children with ASD, and non-verbal subjects. Specific related work is cited in Section 5.2.3 to provide justification for each variable and Section 5.7 in order to assess the current ability to automate video/audio annotation.

5.2.1 General Overview of Direct Observation of Behavior

For many researchers, and many domains, direct observation has been the primary mode of data collection. Before the advent of video systems and digital technology, direct observation was conduced through hand written ethnographies, descriptions and hand sketches of observed behaviors by psychologists who explored man's individual and social behavior, and reports by anthropologists examining the behavior of societies (Bijou et al., 1968; Clifford et al., 1986). As technology has progressed, applications of these same principles have been applied to new domains (Human Factors, Computer Science, CSCW). In addition, the process of gathering observational data has become more operationalized. With celluloid and now digital video, re-watchable sessions can now be annotated and linked with specific behaviors. This allows researchers to quickly refer back to the actual events, rather than rely exclusively on notes and memory (Lee, 1974; Retherford et al., 1993; Rosenblum et al., 2004). To reflect the broad spectrum of disciplines that use video annotation and the study of behavior, many guides and sets of variables have been created. We briefly discuss here the approach of behavioral scientists and researchers in human computer interaction.

5.2.2 Behavioral Sciences

Researchers in the behavioral sciences have studied behavior from a broad range of perspectives: including both diagnostic and therapeutic. Even within different applications, the focus on behavioral observation is quite varied. Some researchers has examined aspects of speech and/or sound production (Koegel et al., 1999; Wetherby et al., 1998; Woods and Wetherby, 2003; Owens, 2007), while others have focused on interaction (with and without researcher/clinician being physically present) (Baskett, 1996; Wetherby et al., 1998). Diagnosis via observation (Lord et al., 2000) and communication skill acquisition (Prizant et al., 1997; Sheinkopf et al., 2000) have also been examined.

There is an interesting parallel between subjects in infant research and older non-verbal subjects in that both populations are non-verbal. In some respects, they present similar levels of verbal competence,

and consequently similar coding challenges. Although our work has a different purpose and is examined in a different context, it shares many of the same critical aspects of behavioral assessment as that of infant research (Segal et al., 1995; Hayne et al., 2000; Luo and Baillargeon, 2005).

Regardless of the specific population, studying the behavior of nonverbal subjects requires a reliable coding system that often is not published. Consequently, investigators in such areas are often forced to "reinvent the wheel."

5.2.3 HCI Research

Computer Scientists, particularly in HCI, and Social Scientists have developed a broad set of coding guidelines for a large array of tasks (Whyte, 1980; Suchman, 1987). However, few of these focus on the evaluation of subjects who are non-communicative. Even fewer address those subjects with ASD. Guidelines exist that have dealt with higher functioning subjects (Piper et al., 2006; Cassell et al., 2007; Gillette et al., 2007; Tartaro and Cassell, 2008) and most, gathered data through subjective (qualitative) observation (Kerr et al., 2002; Michaud and Theberge-Turmel, 2002; Pares et al., 2005).

Although the literature in the HCI and Social Science disciplines is comprehensive and examples of coding guidelines are robust, an established quantitative coding system that addresses the behavior of non-verbal subjects and interventions using computer systems that provide auditory and/or visual feedback does not exist. This chapter addresses this gap by detailing the construction of A^3 coding guideline and the reliability of the variables in an experimental research setting.

5.3 A^3 Coding Guidelines Development Process

From existing areas of research, we drew the most relevant and salient dependent variables in relation to vocalization and engagement to help quantify the interaction of non-verbal subjects and computer-based intervention systems. By drawing theory and experimental findings from 19 sources spanning five disciplines (Special Education, Speech and Hearing Science, Infant Research, Diagnostic Observation, and HCI), we built upon the research of others and grounded our guidelines in the bodies of literature across multiple areas of science. This resulted in a set of 17 momentary metrics and four durational measures, which totaled 21 dependent variables for analyzing the behavior of non-verbal subjects, which we termed "A^3". The name A^3 references the original purpose of studying ASD behavior with computer intervention, though the application of this system is far broader.

To facilitate the refinement and reliability of the variable definitions, four coders were employed to annotate video from the experimental context. All coders were from the Department of Speech and Hearing Science at the University of Illinois at Urbana-Champaign. The first two coders were undergraduate seniors, while the second two were graduates of the undergraduate program (one was pursuing a master's degree in Speech and Hearing Science). All video coders had class experience in phonetic transcription and three of the four had worked as coders on relevant research projects.

5.3.1 Process of Refinement

Beginning in October 2007, researchers began a seven-week iterative design cycle in which the definitions of the 21 metrics for examining behavior were refined. Each week, two coders reviewed the same video from the experimental context and annotated all 21 variables based on the current definitions using the video annotation system VCode(Chapter 4). At the end of each week, coders met with researchers and an examination of agreement was performed using VData (Chapter 4). Discussion of discrepancies resulted in modifications to the variables's descriptions. Each successive week, data were coded in light of the new definitions.

At the conclusion of the seven weeks, two additional coders were recruited due to a change in the availability of the original coders. Though this required training to begin again, the addition of two fresh perspectives ensured that any assumptions about variable definitions made by the first pair of coders (this potential confound is referred to as observer drift — see (Kazdin, 1977)) were revealed and explicitly noted in the guidelines through several weeks of training/guideline-refinement.

This process continued until a point-by-point inter-rater agreement level (Kazdin, 1982) of 85% was reached across all variables in one session. Although an agreement level of 80% is considered acceptable (Kazdin, 1982), we wanted to ensure that the construction of variable definitions was "above" standard. We used a point-by-point agreement calculation with a tolerance of one second (i.e., two events were said to agree if the secondary coder's mark was within 0.5 seconds on either side of the primary coder's mark) for calculation of agreement. See Section 5.6 for a more detailed discussion of agreement calculations employed in this research.

5.3.2 Video Annotation Software

To facilitate the video annotation process, many digital tools have been developed (Burr, 2006; Noldus, 2007; Salt Software LLC, 2007; Kipp, 2007), all of which can be used in conjunction with A^3. With the

onset of these digital systems, software designers can enhance them to aid in the coding process. For our own research, we created a tool called VCode (Figure 4.3) which was presented in Chapter 4. VCode was designed to ease the burden on coders, specifically related to coding schemes such as A^3. Many of the features of VCode and VData are essential to the construction and use of A^3. Specifically, the two modes of interval playback allow for two levels of video analysis; coarse (Interval Playback) and fine (Continuous Interval Playback). Both of these modes have been described in the behavioral observation literature — see (Alberto and Troutman, 2005).

5.4 A^3 Variables

The following section is a detailed description of the dependent variables examined in the A^3 coding guideline. These descriptions focus on the rationale for each variable and the major choices made when constructing the variable definitions. The actual guide (with the specific topographical, physical, or behavioral features for annotation) is presented in Appendix A. In general, we can divide our dependent variables into measures of (a) engagement and (b) vocal behavior of the subject. We begin by highlighting the engagement variables, because engagement is often viewed as a prerequisite to learning and communicative exchange. We follow with a review of the variables based on child vocalization.

Before reviewing the individual variables, it should be noted that one variable, Phonetic Transcription (or marking the phonemes uttered by subjects), was dropped early in the analysis process due to coder feedback. Coders stated that because of poor audio quality and extremely poor articulation by subjects, accurate and reliable phonetic transcription would not be possible. It should also be noted that here we present only justification for each variable, and refer the reader to Appendix A for variable descriptions.

5.4.1 Engagement Variables

With the exception of the metric Time In Chair (Section 5.4.1.5), which was gathered with standard playback, all these metrics were gathered with the Continuous Interval Playback Mode set to three seconds. This mode is ideal for variables that are not discrete (i.e., difficult to specify their exact starting or ending point, or exact duration).

5.4.1.1 Smiling

The variable Smiling was chosen because it is typically associated with pleasure or enjoyment (Field et al., 2001). Although the source of the smile could not always be determined, we hypothesized that a higher rate of smiles would reflect conditions that were generally more enjoyable to the subject.

5.4.1.2 No Face

The No Face variable was used to identify three-second intervals when the child's face could not be seen, and no coding determination could be made as to whether or not a smile occurred. This variable was identified because of a concern that surfaced during the coding process: coders found that when they summarized the data, they had difficulty discerning intervals when no smiles had occurred from those they were unable to code. Although its accuracy is reported, this variable was not directly used in the analysis. Rather its agreement was useful for demonstrating that coders were "on the same page," and allowed agreement calculations for Smiling, which is dependent upon being able to see the face. The absence of both Smiling and No Face marks are an indication that the child was not smiling.

5.4.1.3 Oriented at Screen

In order to assess visual attention to content, we created an "orientation arc" for the evaluation of the child's gaze. See (Baskett, 1996; Field et al., 2001) for others who opertaionalized this procedure. If gaze was directed within this arc within the 3-second interval, the subject was considered to be Oriented at Screen. The arc's width (90ą) was used to accommodate the behavior in which children with ASD use peripheral vision as primary visual input (Howlin, 1986). See Appendix A for illustration/diagram of the arc.

5.4.1.4 Auditory Focus

Much like Oriented to Screen, the Auditory Focus variable was used to assess auditory attention. Unlike visual attention, which has a more observable physical indicator, auditory attention must be observed indirectly. Auditory Focus was observed via changes in subject proximity to (moving closer) and physical contact with the speaker which is a possible surrogate for inferring interest in computer audio. In addition, Auditory Focus was also annotated if the subject oriented to the screen/speaker after a new sound was made by the computer (Clifton et al., 1999; McCalla and Clifton, 1999).

<table>
<tr><td colspan="6">Child's Sounds</td></tr>
<tr><td colspan="2">Non-Speech</td><td colspan="4">Speech Like</td></tr>
<tr><td>Laughter</td><td>Other</td><td colspan="2">Imitative</td><td colspan="2">Speech Like</td></tr>
<tr><td colspan="2"></td><td>Immediate</td><td>Delayed</td><td>Oriented at Screen</td><td>Not Oriented</td></tr>
<tr><td colspan="4"></td><td>Immediate</td><td>Delayed</td><td>Immediate</td><td>Delayed</td></tr>
</table>

Figure 5.1: *Child Sound Decision Hierarchy*

5.4.1.5 Time in Chair

To assess the willingness to attend to computer stimuli, we coded the duration a child would spend in his/her chair (Sajwaj et al., 1972; Lovaas, 2003). We hypothesized that increased time sitting was a proxy for engagement with the computer.

5.4.2 Verbal Variables

Verbal metrics were collected to examine vocalizations during the experimental and control conditions. Coding of vocalizations was facilitated through use of a decision tree, which is incorporated in the full coding guide (Appendix A). Figure 5.1 presents the child sound decision tree in isolation as a hierarchical set. From this perspective, researchers can better understand the relationships between different variables. With the exception of Turn Taking (Section 5.4.2.7), the following sub-sections are presented hierarchically at decision points in the tree rather than one for each variable. These variables were assessed with Standard Playback. It should be noted that we present only justification for each variable, and refer the reader to Appendix A for variable descriptions.

5.4.2.1 Child's Sound (Speech vs. Non-Speech)

The most basic question coders had to address was whether or not a sound was considered "speech-like." Specifically, we define a Speech-Like sound as one that could be phonetically transcribed. This decision point attempts to screen sounds that have the potential to lead to conventional speech and those that may be related to ticks, breathing, self-stimulatory behavior, or other forms of expression that are not used in speech production (laughing, screaming, etc) (Woods and Wetherby, 2003).

5.4.2.2 Non-Speech Sounds (Laugher vs. other)

Though the range of Non-Speech Vocalizations is large, we asked coders to distinguish Laughter as another means to examine children's pleasure and engagement during the activity (Gena et al., 1996). (We realize that this requires an inference that may not be accurate for many children with ASD.) In addition, we wanted to differentiate Laughter from vocal self-stimulatory behaviors. Compared to other children with developmental delays, those with ASD tend to produce more non-speech sounds (Sheinkopf et al., 2000). By annotating Non-Speech Vocalizations, we also hoped to examine the impact of the external stimuli on their non-speech vocalization and in comparison to their Speech-Like Vocalization.

5.4.2.3 Speech-Like Sounds (Imitative vs. Spontaneous)

A critical distinction made in studying the communicative behavior of children with special needs is between sound production that is Imitative (repeating a sound previously heard) or Spontaneous (without an immediate model) (Halle, 1987). Using this distinction, we hoped to explore the nature of speech-like sounds produced and if there is a direct relationship between what is prompted (human/computer) and what is vocalized.

5.4.2.4 Imitative Sounds (Immediate vs Delayed)

To explore the imitative sounds produced by subjects, we divided them into those which occur immediately after a source (within five seconds) and those that occur after a more prolonged time (Prizant et al., 1997). This distinction is particularly relevant for children with autism due to echolalic tendencies (Rapin and Dunn, 1997). Theory suggests that words/sounds in delayed imitation are stored outside of the subject's short-term or working memory (Gathercole and Baddeley, 1993; Bradford-Heit and Dodd, 1998).

5.4.2.5 Spontaneous Sounds (Orientation to Screen)

While imitative sounds are, by definition, based on audio stimuli, we wanted to delve deeper into spontaneous sounds, and their relationship to screen orientation. Eye gaze is indicative of engagement. When paired with vocalization, it is a key communicative development (e.g., (Baskett, 1996; Wetherby et al., 1998)). For each spontaneous sound produced, we explored whether or not that sound was made

while oriented to the screen. This allowed us to examine the direct relationship between spontaneous sound production and orientation.

5.4.2.6 Spontaneous Sounds (Immediate vs. Delayed)

Much like imitative sounds, we wanted to understand if there was any correlation between spontaneous sound production and auditory stimuli. To explore this relationship, we asked coders to mark spontaneous sounds that were made within five seconds (immediate) of a source sound, and those made after a longer period of time (delayed).

5.4.2.7 Turn Taking

An important skill in oral communication is that of turn taking, or waiting for others to finish their utterance before vocalizing (Wetherby et al., 1998; Owens, 2007). With all speech-like sounds, we asked coders to determine if the subject waited for the source (be it a researcher or computer-generated sound) to "finish" their sound production. In other words, did the child wait for his/her turn to talk (or not interrupt).

5.4.3 Other Metrics

Previous sections dealt with variables related to Engagement and Verbal Behavior. Two non-verbal engagement variables useful in analyzing subject behavior are presented here.

5.4.3.1 BIGmack Switch

The BIGmackł Switch (AbleNet, 2008) is an assistive technology device that plays a pre-recorded sound for individuals with speech and language delays (Reichle et al., 2002). With one subject, who was suspected of having limited motor control of his vocalizations, this device was used to simplify the task of producing sounds.

5.4.3.2 Non-Child Audio

These data points served two purposes. Primarily, they were collected to help clean data logged on the computer by marking sounds (other than those made by the child) in the video that interfered with

automatic data gathering. A second purpose was to familiarize coders with the video they were about to watch without forcing them to annotate a very complex variable. Because Non-Child Audio was coded as a first pass, by itself, coders were required to watch the entire video once, before examining more specific details.

5.5 A^3's Four Pass System

The actual annotation process was divided into four passes, each of which asked coders to focus on a specific category of dependent variables while they watched a video in its entirety. Since many of the different variables required different view modes, this pass breakdown not only aided the examination of the data, but also was optimal for the annotation process. The specific variables for each pass can be seen in Appendix A.

5.5.1 Pass 1 – Data Cleaning

In addition to providing visual and audio information to subjects, computers have the ability to perform real-time analysis of audio (from a microphone) and visual (from an attached camera) data. To draw corollaries between changes in the state of the computer system and the behavior of subjects, HCI systems often log a robust set of data. Though these data can be rich and useful in behavior and interaction analysis, this data set can often be muddied by non-valid audio data (audio input that came from non-subject sources). In other words, sound from a researcher, an object falling, or the subject hitting the computer/table can appear in the logged data. The first pass focuses on reducing the "noise" from the audio data to help filter the logged information.

5.5.2 Pass 2 – Attentiveness and Engagement

The second pass is a collection of variables that can be used to assess the engagement, response, and interaction of the subject during presentation of the computer's audio and visual feedback. The data are collected using a Continuous Interval Playback.

Variable	% Agreement
Non Child Audio	75.96
Duration: Non child audio	82.39
Smiling	90.71
No Face	86.01
Oriented at Screen	98.09
Auditory Focus	87.37
Laugh	63.89
Duration: Laugh	91.30
Non Speech Vocalization	83.86
Speech-like Vocalization	90.50
Duration: Speech like vocalization	92.84
Turn taking	81.60
Immediate + Screen + Spontaneous	79.30
Delayed + Screen + Spontaneous	79.45
Immediate + Spontaneous	78.31
Delayed + Spontaneous	84.62
Delayed Imitation	*None Recorded*
Immediate Imitation	85.71
Time in Chair	88.64
Duration: Time in Chair	82.05
BIGMack Switch	87.50

Table 5.1: *Point-by Point Percent Agreement*

5.5.3 Pass 3 – Vocalization Analysis

The heart of the A^3 coding guidelines is a detailed analysis of the subject's vocalizations, specifically focusing on those that are considered to be "speech like," or have the potential for contributing to meaningful speech.

5.5.4 Pass 4 – Time in Chair

The fourth pass examines the time a subject spends in the chair. Though not all subjects with developmental disorders are willing to sit for an extended amount of time, analysis of this behavior can be used to examine participation with the computer system.

Variable	% Agreement
Immediate Spontaneous	83.19
Delayed Spontaneous	80.57
All Spontaneous	84.62

Table 5.2: *Combined data from the four spontaneous speech-like vocalization variables*

5.6 A^3 Reliability & Efficacy

By situating the creation of A^3 in an experimental context (Chapter 6: Spoken Impact Project (SIP)), researchers were able to utilize the collected video data to test reliability and efficacy of the A^3 system. During data collection using VCode, coders were asked to annotate all 21 variables. Of the 268 trials, 11% were checked for reliability (132 minutes of video footage). Not all trials were equal in length, nor did all variables occur equally in all sessions. As a result, we examined agreement values across randomly sampled sessions. We present reliability calculations, coder's feedback, and a demonstration of efficacy of A^3 in the experimental setting.

5.6.1 Point-by-Point Agreement

Using point-by-point agreement method (Kazdin, 1982), overall agreement between coders across all variables was 88%. Agreement was defined using a conservative tolerance of one second (two events were said to agree if secondary coder's mark was within 0.5 seconds on either side of primary coder's mark). Percent agreements are presented in Table 5.1. Five variables had an agreement above 90%, 11 above 85%, and 15 were above 80%. However, five variables fell below the 80% benchmark for reliable data collection. The sub-sections of 5.6.2 below address these discrepancies.

5.6.1.1 Robustness of Agreement Calculations

To observe how agreement would change with an increase in the timing tolerance, we increased timing tolerance from 1.0 to 5.0 seconds at 0.1 second intervals. Surprisingly, the number of matched points changed only when the tolerance was increased to 1.5 seconds (observers' marks were said to agree if the secondary mark was within 0.75 seconds on either side of the primary coder's mark). Of the few variables whose reliability increased, the change resulted in a gain of at most 0.5%. This suggests that the coders were likely accurate in the placement of their marks. Moreover, we can surmise that if there were a lack of agreement in the data, it was likely due to a disagreement of what was coded and not due to ambiguity regarding whether an event had occurred.

Variable	Kappa
Auditory Focus	0.64
Oriented at Screen	0.84
No-Face	0.77
Smiles vs. No Smiles	0.63

Table 5.3: *Kappa Statistic for variables coded using interval playback (set to 3 seconds)*

5.6.2 Discussion of Variables with Low Agreement Scores

5.6.2.1 Laughter

The variable with lower inter-rater agreement values is Laughter, with an agreement of 63.89%. However, upon further inspection, we noticed that it also had a low frequency of occurrence. Lower levels of agreement are often achieved on low-rate behaviors because there are fewer opportunities to observe the target variable (Birkimer and Brown, 1979).

Coders also mentioned difficulty in distinguishing Laughter from Speech-Like Vocalizations and Non-Speech Vocalizations. Often coders found that vocalizations may have been laughter-like, but matching positive affect with the vocalization was difficult. Perhaps this difficulty is exacerbated by the differences in affect expression that characterizes ASD (Wetherby et al., 1998). Despite this characteristic, reliability of Laughter may be improved through increased microphone quality. An explicit plan to strengthen the reliability should be considered.

5.6.2.2 Spontaneous Speech-Like Vocalizations

The sub-divisions within Spontaneous Speech-Like Vocalizations (with and without orientation at screen and delayed vs. immediate) also resulted in low agreement between coders (78.31% to 84.62% with a mean of 80.42%). To explore the effect of sub-division on reliability, we combined data across variables (Table 5.2). To combine two variables, we treated all marks for both variables the same and re-calculated agreement. This analysis can suggest at what level of granularity (distinction between variations of a variable) these variables can be reliably coded. When we eliminated the distinction between sounds made while the subject looked at the screen versus those made when looking away, reliability improved to 80%. It appears that such small distinctions may have been too fine-grained to code accurately unless multiple camera angels are available. Otherwise, we recommend only differentiating Immediate and Delayed Spontaneous Speech.

Upon further examination of the Spontaneous Speech-Like data, we discovered the potential for double-counting disagreements. Every Speech-Like Vocalization was coded as either Spontaneous or Imitative. However, every disagreement in Speech-Like Vocalization is a guaranteed disagreement for the Spontaneous/Imitative distinction. Thus, our lower agreement values may have been a direct result of "double counting."

From this analysis, we believe that the best approach for implementation of A^3 is to code all levels of Spontaneous Speech-Like Vocalization that are of theoretical interest, and consider the possibility that distinctions between Immediate and Delayed Imitations may need to be collapsed to achieve reliability. To analyze the effects of screen attention and vocalization, we suggest performing a post-hoc analysis between Oriented at Screen and Speech-Like Vocalization. Although this may not link each specific vocalization to the visual display, a trend should be apparent and extrapolation should be possible.

5.6.2.3 Non-Child Audio

The second variable that had less than 80% agreement was Non-Child Audio. One plausible explanation for the poor agreement in this variable may be due to the poor quality of the audio recording. Some coders were better able to hear quieter sounds and, thus, would mark them, while other coders could not hear these sounds and they would escape notation. As a result, there was a discrepancy between the coders. We propose an audio "threshold" be set to differentiate between sounds to code and sounds not to code through the use of a low-pass filter. With the addition of a low-pass filter (either to be employed by the computer or presented as a visualization inline with the coding timeline), we believe we can surpass the 80% agreement level as recommended by Kazdin (Kazdin, 1982).

5.6.3 Kappa Statistics

Most variables were coded on an infinite timeline. In other words, marks could be placed at almost any location during the duration of the video. As a result, the probability of a chance agreement was extremely low, and thus reduced the need and applicability of Cohen's Kappa (Kazdin, 1982). However, for the variables collected using the Continuous Interval Playback, we calculated Kappa analysis because the video was annotated in discrete, binary segments (set to 3 seconds) of required observations, and thus subject to chance observation.

The Kappa statistics calculated from the data suggest a high level of agreement (Table 5.3). Kappas ranged from 0.64 (Good) to 0.84 (Very Good). Our interpretation of agreement follows from that of

Altman and Byrt (Altman, 1991; Byrt, 1996). These findings add further support to the findings of the point-by-point agreement analysis.

For Smiling and No Face calculations, we used a 2-tier evaluation metric similar to that used by Reid et al. (Reid et al., 1985). The first Kappa accounts for the level of observer agreement on whether they could make a judgment about a subject's smiles (could they see the subject's face – No Face). We then eliminated all of the intervals in which either coder marked No Face implying that they could not assess whether the subject was or was not smiling because the assessment of agreement on Smiling depended on both observers being able to see the subject's face. The intervals, in which both observers coded the subject as either smiling or not, were examined for agreement. In other words, coding of smiles depended upon both observers agreeing that they could see the subject's face.

5.6.4 Efficacy of A^3 in Experimental Context

In Chapter 6: Spoken Impact Project (SIP), we performed an analysis of the experimental context using the A^3 coding guidelines. This chapter examined the impact of different forms and modalities of computer generated feedback on the Spontaneous Speech-Like Vocalizations of five non-verbal participants with ASD. Results demonstrated statistically significant effects of different feedback systems on the subject's vocalization and highlighted their individualized preferences. Most notably, analyses using the A^3 coding guidelines revealed statistically significant differences in vocalization rate based on the modality of the feedback with three children producing more Spontaneous Speech-Like Vocalizations in the presence of audio feedback, and one child responding more consistently to visual feedback. One child's Spontaneous Speech-Like Vocalization rate was unaffected by the type of feedback systems. One of the three children who responded significantly in the presence of audio feedback also demonstrated a statistically significant response to conditions with both audio and visual feedback. These results illustrate the capacity of the A^3 coding guidelines to quantitatively assess vocalization, but they also support qualitative observations made by researchers.

During the experimental period, researchers made qualitative observations of participants' behavior, their engagement, and which forms/modalities of feedback produced better responses. Upon analysis of the variables collected using A^3, statistical findings mirrored most qualitative observations made by researchers. This suggest that the A^3 coding guidelines can quantitatively capture qualitatively observed behaviors.

This investigation supported the effectiveness of A^3 for quantitatively assessing vocal behaviors in nonverbal participants (see Chapters 6, 7 and 8). Currently, we are expanding our data analysis in the

experimental context to further assess the impact of computer-based feedback systems on measures of engagement, as well as the differential impact on speech-like versus non-speech vocalizations in the same participants.

5.6.5 Coder Feedback

By the last video, coders spent approximately 20 minutes to annotate one minute of video footage. This represents a significant decrease from the initial 40 minutes per one minute of video footage (self reported by coders). Coders felt that the majority of their time was spent on the third pass of the annotation system, due to the complexity and scrutiny required to differentiate among different types of vocalization. This difficulty was exacerbated due to the poor articulation of the the population used in the experimental context and the poor quality of the video camera's microphone. Improvement of microphone quality could help improve accurate labeling of the vocalizations.

In further discussion, coders mentioned some confusion over the No Face variable, specifically in the boundary condition when the child has part of his or her face covered. Feedback from coders included specific requests for a more explicit definition of the features that must be seen to justify annotating No Face. For future experiments, we propose specifying features of the face (e.g., lips, cheeks) that most clearly provide access to determination of subject: re affect.

5.7 Automation and A^3

Although coding time was reduced from a self-reported 40 minutes/1 minute of video to 20 minutes/minute of video, coding remains time consuming. For researchers, this impacts the time required to gather data, as well as delays data analysis and the ability to progress in the research itself. Likewise, clinicians need to reduce time demands in order to provide clinical utility (Perlman, 2008). As VCode was developed to assist in the coding process, further computerized systems can be developed to reduce the burden on coders and the time required to annotate a video. Specifically, through the use of computer automation, the task of coders may be substantially reduced if not eliminated. The addition of computer-based automation has the potential to not only increase the speed at which video-annotation can be accomplished, but also to improve the quality of the data gathered for the experimental evaluation (Burr, 2006).

This section details how automation, applied to A^3 could be used to enhance the collection of data. In order to uncover ways the technology could be used to automate the coding process, we examined the existing body of literature and the current state of computer vision and voice recognition software. By examining what technology is capable of today, we are able to assess the degree to which computers could be used by coders and researchers to augment the video annotation process. While research is continuing to make strides in computer vision and speech recognition, we focus our discussion on the state-of-the-art and what can be implemented today, rather than on where technology may be in the future.

5.7.1 Degrees of Automation

With such a complex set of variables and behaviors incorporated in the A^3 guideline, the varying forms of assistance technology can provide are prodigious. We therefore break down each form of technological assistance into one of four categories:

REPLACEMENT: When a coder can be replaced entirely by an automated process for collection of a specific variable. This requires the automated system to have a low rate of false positives and low rate of false negatives. By using a Replacement system, coders need not manually annotate a video for a specific behavior. This form of assistance requires technology to achieve nearly the same agreement as that of a coder. Though this is by far the most demanding on the technological support system, it does greatly reduce the burden on the coder and speeds up time to annotate. For this automation, incorporation with the existing coding practices is not required. Rather, researchers can omit manual annotation of the specific variable.

REVIEW: When a coder is only needed to double check data annotated by an automated process. This requires the automated system to have a low rate of false negatives and a moderate to low rate of false positives. Systems that use the Review model relatively accurately mark when a specific behavior occurs in a video. However, due to a modest rate of false positives, a coder must check over each annotation to assess its validity. This substantially reduces time required to annotate, because coders are not required to watch a video in its entirety. Rather, they jump from mark to mark. Low false negatives are a requirement, in that, coders will not be watching non-marked footage. Hence, missed behavior will not be marked. For this automation, incorporation with existing practices can be achieved by creating annotations that exist in the coding environment (e.g., as actual marks on a timeline) or as a secondary display of information (e.g., as in the graphical data shown on along side the annotation timeline).

QUICK INDEX: When the automated process can help guide coders where to "look" for a behavior, but not provide actual annotations for use in data collection. This requires the automated system to have a low rate of false negatives and can accommodate a high rate of false positives. Quick Index automated systems point out time spans of interest to coders, allowing them to skip areas of inactivity. By maintaining a low rate of false negatives, coders can be assured that the area they are skipping is of little or no value to the current variable. These Quick Index systems can be used when a computer can detect that "something" is occurring, but are unable to identify or categorize the exact behavior that is being demonstrated. For this automation, incorporation with existing practices can be achieved by creating a video stream whose colors change to alert coders that a certain condition is occurring (e.g., as a secondary video stream) or as a secondary display of information (e.g., as a graphical data shown along side the annotation timeline).

MANUAL: When the automation process has a high rate of false positives and a high rate of false negatives, coders must annotate without any assistance. Manual annotations require time and the full concentration of a coder. While these forms of annotations are time demanding, with proper training and regular agreement checks, the accuracy of these annotations (as demonstrated in this chapter) can be reliably used for data collection.

5.7.2 Variable by Variable Automation

This section presents an analysis of each variable and to what degree current technology can facilitate automated annotation. Based on this analysis of existing technology, researchers can design systems to automate their own use of the A^3 annotation guidelines, thereby improving accuracy while reducing coding time. Further, researchers can augment existing open source video annotation software, such as VCode (Chapter 4), to incorporate these forms of automatic video annotation. We break down the variables at key decision points paralleling Section 5.4. A summary of variables and their degree of automation is presented in Table 5.4 at the end of this section.

5.7.2.1 Engagement Behavior Variables

Behavior detection in video is a growing field of research in areas such as scene based automatic annotation (Ponceleon and Srinivasan, 2001; Chen et al., 2002), automatic event logging (Banerjee et al., 2004), and object of focus identification (Bertini et al., 2004). We focus here on systems related to A^3 Engagement Variables.

5.7.2.1.1 Smiling Because positive affect is part of the definition of Smiling, detection of emotion is critical towards automating the process of smile detection. A growing area of HCI research has been focusing on bringing emotion and emotion detection into day-to-day computing (Crane et al., 2007). Within this subset of work, emotion detection has been a critical component (Zacks et al., 2001; El Kaliouby and Robinson, 2004; Wang et al., 2007; Zeng et al., 2007). In traditional HCI, this detection can be used to adapt computer systems to respond more appropriately to human reaction (Setlur and Gooch, 2004). In addition, a growing set of work in facial feature detection has been focused on disability research (Kaliouby and Teeters, 2007; Madsen et al., 2008). Other researchers have examined detection of emotion/arousal with other non-video metrics (Chang and Ma, 2002; Jones and Troen, 2007). This work, combined with research on automatic smile detection (Valstar et al., 2007; Valstar et al., 2006) can provide a Review based automation system. With poorer accuracy, these techniques could still provide a time saving gain of a Quick Index system, simply by identifying "happy" emotional points.

5.7.2.1.2 No Face Research on face detection has long been a major focus of computer vision systems (Zhao et al., 2003). Given this robust set of data, computer systems could relatively easily note moments when no face is visible. However, the definition of No Face only is marked when the face is occluded enough to prevent analysis of a smile. Work in machine learning (Osadchy et al., 2007) allows for facial orientation analysis, which could further be used to determine No Face. Working with research on smile detection (Section 5.7.2.1.1), systems could conceivably be constructed to determine when a smile was not present, however, that does not necessarily mean the face was occluded enough. In conclusion, technology has great potential here, however, current algorithms are not robust enough to facilitate Replacement or Review systems. At best, a Quick Index system could be constructed to alert coders to the presence of a face, and thus require coders to mark when a No Face truly occurred.

5.7.2.1.3 Oriented at Screen In addition to using face detection techniques (Zhao et al., 2003), eye-tracking technology (Jacob, 1991) can be used to assess when subjects are looking at a screen (Wang et al., 2007). Further, work in machine learning (Osadchy et al., 2007) can also utilize facial pose and orientation to determine looking direction. Because A^3's guideline examines facial orientation within an arc, we believe that existing technology can provide a fairly robust Review automation system, for annotating when subjects are oriented. However, it would still be necessary for coders to review these marks to reduce the false positive rates and determine when faces detected are not those of the subject.

5.7.2.1.4 Auditory Focus Auditory Focus is a complex relational analysis between audio produced by the computer and the child's orientation to the computer. However, these components, when broken down, can be detected relatively accurately by a computer system. Section 5.7.2.1.3 detailed the detection of subject orientation to screen. That combined with logging of computer-produced sound occurrences, can create a Reviewable automation system, requiring coders only to validate the behavior. For the more mundane physical changes in proximity to an audio speaker or making physical contact with a speaker, using computer vision becomes increasingly difficult. At best, computer vision systems can detect motion (Zhao et al., 2003; Xiao et al., 2008). Setting a certain degree of motion, systems could be automated to note locations for Quick Index.

5.7.2.1.5 Time in Chair Time in Chair could be easily determined by a series of pressure sensors, detecting the presence/absence of the weight of the subject in the chair. The pressure sensors would have to be calibrated to detect full sitting, but not those when child is not "sitting" (e.g., one leg on chair and one leg standing), resulting in Automatic automation.

5.7.2.2 Verbal Variables

Voice detection and analysis is a rapidly evolving field of research covering dialogue transcription (Johnson, 2007; Salt Software LLC, 2007; Studiocode Business Group, 2007) and Voice-Computer Interaction (IBM, 2005; Nuance Communications, 2008). A large hurdle when identifying and classifying speech comes from speaker identification (Nishida and Ariki, 1998; Kwon and Narayanan, 2004). However, the largest burden still comes from identification of what is being said. Though this is a robust field of research, it does require a large degree of training data for each subject. Much work has focused on identifying what is being said for subjects with good diction and vocalization. Thus it is unclear how well it would function with non-verbal sound production or vocalization of individuals with speech impairments. As a result, much of this classification may occur largely Manually. However, by at least noting when sound does occur (though volume analysis), coders may be able to use Quick Index from this degree of automation.

5.7.2.2.1 Child's Sound (Speech vs. Non-Speech vs. Non-Child-Audio) Identifying the difference between speech and non-speech vocalizations would require a robust speech detection engine to make this distinction. While much of the existing literature has focused on identifying words and letters in speech (IBM, 2005; Nuance Communications, 2008), very little has been done to examine

those sounds produced that are not speech-like. Even more complex is identification of sounds that are not speech at all, rather other miscellaneous sounds in the environment. Some more recent research in the area of speech separation has begun examining and differentiating different sources of sounds coming from a solo microphone (Bach and Jordan, 2006; Olsson and Hansen, 2006). Without further work, automation is limited to identifying when sound occurs in a video. This identification, of Quick Index automation, should decrease annotation time. However, the burden of identification and classification is on the shoulders of coders.

Additional automation can be done to detect the "breaks" between vocalization, and should aid in the breaking up of vocalizations. However, when there are simultaneous sounds (subject vocalizing while something is occurring in the room), meaningful (and separate) vocalizations may be linked together due to background noise. Thus, this breakup can be helpful, but at most, would provide a Reviewable automation.

5.7.2.2.2 Non-Speech Sounds (Laugher vs. other) Research on detection of laugher has mainly focused on audio processing (Melder et al., 2007). Some new work has added the additional analysis of vision and audio processing to detect laughter (Petridis and Pantic, 2008). This new work demonstrates an 80% precession rating. This vein of research, in concert with presence of audio and emotion detection (Section 5.7.2.1.1), has the potential to act as a Quick Index or possibly a Review system (with suitable improvement in the technology). Detection of other types of non-speech sounds will probably remain at a very course level of Quick Index automation, simply based on the presence of audio in a video. Though this will reduce time required for video analysis by allowing coders to skip over non-relevant portions of video, current speech recognition systems are not able to differentiate between generic "non-speech" sounds and those that are Speech-Like.

5.7.2.2.3 Speech-Like Sounds (Imitative vs. Spontaneous) The distinction between Imitative and Spontaneous can be automated to a small degree though analysis of the phonemes via speech detection (IBM, 2005; Nuance Communications, 2008) and syllable detection (Howitt, 2000). However, though this distinction may be somewhat effective, gauging appropriate start/stop of vocalizations falls under the same accuracy concerns in Section 5.7.2.2.1. Further, not all sounds come from a source that can be imitative. Thus, difficulties with speaker identification (Nishida and Ariki, 1998; Kwon and Narayanan, 2004) are also problematic. The resulting automation, at best, is Quick Index identification, requiring a heavy degree of analysis, and examination of spaces that are both marked

and non-marked. Until technology better supports a host of concerns outlined here, it is most like a Manual annotation system.

5.7.2.2.4 Speech-Like Sounds (Immediate vs. Delayed) In general, computers should be able to differentiate between sounds with a short delay and sounds with a long delay. This can be accomplished through simple analysis of the presence of sound and the time between sounds. However, there is a great deal of review necessary due to concerns outlined in the above sections related to relevant sounds and appropriate division of overlapping sounds. Yet this additional automation should greatly help coders classify a majority of the sounds. They will still need to re-listen to each vocalization to ensure it is divided appropriately. This level of Review automation is quite plausible and easily implemented.

5.7.2.2.5 Vocalization and Orientation to Screen Given the state of facial orientation software discussed in Section 5.4.1.3, computer systems could go through a set of annotations made by coders and mark those vocalizations that were made while oriented at the screen given the A^3 guideline. Post automation, coders can Review those marks, just to verify the occurrence of the orientation. Given the rapid improvement of computer vision technology, this degree of automation may become a Replacement for manual annotation.

5.7.2.2.6 Turn Taking Much like the problems from speaker identification (Nishida and Ariki, 1998; Kwon and Narayanan, 2004), work analyzing turn taking requires identification of two sounds occurring simultaneously. For those experimental set ups that use computer-based audio, a simple logging system would allow for automation between the existence of sound in the video that occur at the same time that the computer logs generating the sound. However, without better source identification, the limitation of the automation is a set of Quick Index automated location detection, in which, coders can review events when there was video sound, and sound in the video.

However, for turn taking conditions with a human subject, automation is dependent upon technological advances in speaker identification systems. The main existing solution requires each subject to have a microphone, and use subtraction to determine overlap. Until the time of accurate and robust systems, coders must Manually annotate multi-person turn taking.

Variable	Automatic	Review	Quick Index	Manual
Non Child Audio		○	●	●
Smiling		◑	●	●
No Face			◑	●
Oriented at Screen		●	●	●
Auditory Focus		◑	●	●
Laugh		○	●	●
Non Speech Vocalization		○	●	●
Speech Like Vocalization			◑	●
Turn Taking		◑	●	●
Immediate vs. Spontaneous			○	●
Immediate vs. Delayed		◑	●	●
Orientation + Vocalization	○	◑	●	●
Time in Chair	●	●	●	●
BIGMack Switch	●	●	●	●

Table 5.4: *Degree of Automation by Variable*
● *Possible/Usable* ◑ *Probable* ○ *Potential*

5.7.2.3 BIGmack Switch

By attaching a pressure sensor to a BIGmack Switch, computers can log this behavior, and Replace coders. Pressure sensors can also be applied to other external therapeutic and communicative devices (e.g., GoTalk 20+ (Attainment Company, 2008)) to accurately log usage.

5.7.3 Implications of Automation

Based on the analysis of the existing technology, computers can greatly augment the video annotation process through automation. The creation or modification of automated systems can improve accuracy and reduce coding time of the A^3 annotation guidelines. We provide a summary table listing each variable with the degree of automation available (Table 5.4). Overall, computer based automation can aid coders by highlighting areas of meaningful activity, reducing the time required to view video segments that have no relevant behavior. For eight variables, automation is accurate enough to only require coders to validate the automated marks, thereby reducing time required and greatly improving accuracy. The resulting use of these techniques may be through individual monitoring systems that log specific behavior (e.g. checking visual orientation) a plug-in to a system like VCode that analyzes the video/audio streams, or a completely new package that provides multiple forms of automation for clinicians and research, aggregating and logging many data points. Technology is currently not

advanced enough to fully remove coders from the video annotation process. However, as video and sound analysis techniques continue to improve, so will the burden placed on manual coding.

5.8 Forms of Use

Though demonstration of A^3 came in the context of coding variables from video of children with ASD interacting with technology, we strongly believe that the application of A^3 extends well beyond this particular situation.

5.8.1 Application External to ASD

Because this guideline focuses on non-verbal subjects interacting with technology, we believe that A^3 can be applied to other areas of HCI research that target non-verbal subjects. This could include infants, non verbal subjects with verbal apraxia, autism, or other severe and profound disabilities. Because our system also focuses on computer feedback and categories of vocalizations, A^3 could also be used in situations where subjects use augmentative and alternative communication (AAC) speech generation devices (e.g., Go-Talk(Attainment Company, 2008)) (Reichle et al., 2002).

Beyond expanding A^3 to multiple disorders, it can also be used for data driven clinical assessments and intervention. During speech and behavior therapy, researchers often incorporate different technological tools. This ranges from chapter constructions to AAC devices and computerized speech feedback systems (the type of application used in the experimental context). In the therapeutic context, clinicians also must be able to assess progress made by subjects. A^3 allows practitioners to assess changes in behavior, as well as the subjects' engagement with technological devices used in therapy.

5.8.2 Use of the A^3 Guidelines

Though A^3 does not cover every possible dependent variable for video analysis, it does provide a robust library of features to annotate. By understanding the interaction between subjects and computers through video annotation and A^3, researchers can evaluate the software intervention itself. However, not all features of A^3 may be applicable, or worth analyzing for every experiment. Thus, A^3 is designed to act as a source for dependent variable selection. When designing an experiment, researchers can select the most applicable set, definition, and justification of variables from the guideline, and use that

sub set in their own research. By utilizing the four-pass system, researchers can extract passes that directly target the type of analysis they wish to perform.

Variables are not limited to simple frequency counts. They can be presented as rates of behavior (e.g., vocalizations/10 seconds) or ratios (e.g., Speech-Like Vocalizations: Non-Speech Vocalizations). By analyzing ratios, researchers can assess the influence on multiple behaviors concurrently. A^3 presents multiple related perspectives on many types of variables (e.g., Speech-Like Vocalizations vs. Non-Speech Vocalizations). We believe that the strength of A^3 is its flexibility allowing researchers to examine their own set of variables in the most efficient way, using a set of operationalized dependent variables.

5.8.3 Coding Time

Though the coding time for all 21 metrics was 20 minutes/1 minute of video, this was the most intensive form of video annotation required by the A^3 system. As researchers refine their selection of variables, the time required to annotate video will be greatly reduced. In addition, improvements in microphone quality will also improve speed of analysis in that coders will not need to repeat as many segments of video for clarification. Time accelerators based on computer-based automation can be found in Section 5.7.

5.9 Conclusion and Future Work

Research with subjects who are non-verbal poses challenges for researchers in terms of creating educational solutions and testing computer-based interventions. Without an existing and unified set of metrics for assessing these subjects' interaction with technology, a reliable and standardized method for comparing and contrasting the performance of different systems does not exist. This chapter presents A^3 (Annotation for ASD Analysis), a collection of dependent variables for video annotation that can be used to assess the interaction between non-verbal subjects and computer based interventions. As researchers, it is incumbent upon us to faithfully describe our scheme and its relationship to other tools, and provide data to support its reliability in the field.

In addition to providing the operationalized coding process and detailed justification grounded in existing literature, this chapter presents an analysis of existing technology and how it can be used to automate the annotation of many of the variables. The degree of automation can reduce the time

required for coders to annotate, in addition to increasing accuracy, and reducing costs for researchers. This analysis is grounded in the capabilities of existing technology, based upon a detailed examination of literature. Lastly, we present a detailed description of where and how the A^3 coding guidelines can be used. For future experiments, we hope to design tools to further aid in the automation of A^3 variables, as well as improve the description of some of the variables based on of the lessons learned. Overall, A^3 provides researchers a unified source of operationalized variables that can be used to assess the interaction between nonverbal subjects and computer-based intervention systems.

Chapter 6

Spoken Impact Project (SIP)

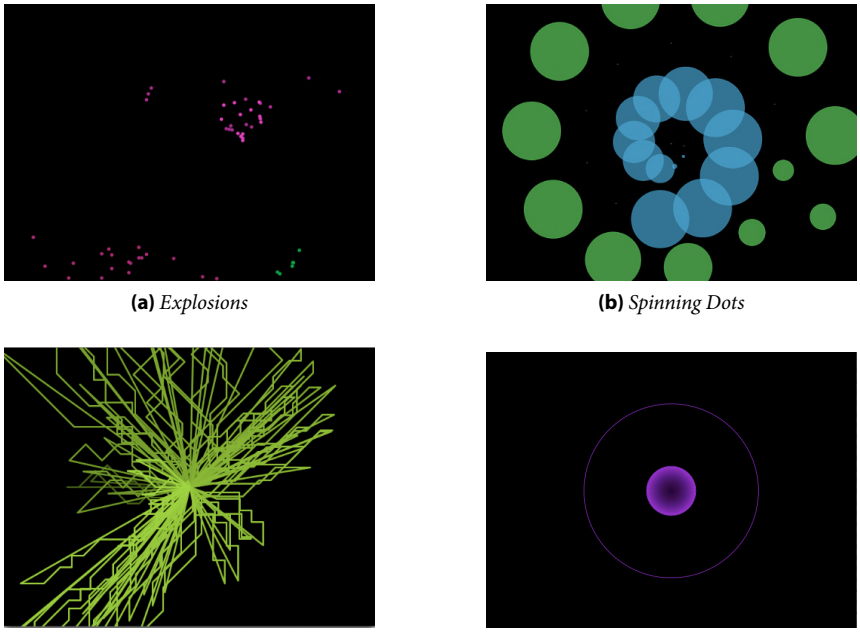
One hallmark difficulty of children with Autism Spectrum Disorder (ASD) centers on communication and speech. Research into computer visualizations of voice has been shown to influence conversational patterns and allow users to reflect upon their speech. This thesis presents the Spoken Impact Project (SIP) which examines the effect of audio and visual feedback on vocalizations in low-functioning children with ASD by providing them with additional means of understanding and exploring their voice. This research spans over 12 months, including the creation of multiple software packages and detailed analysis of more than 20 hours of experimental video. SIP demonstrates the potential of computer generated audio and visual feedback to shape vocalizations of children with ASD.

6.1 Spoken Impact Project Software (SIPS)

During three months (Summer 2007), researchers designed the Spoken Impact Project Software (SIPS) package in Java using the Processing Library (Fry and Reas, 2007). SIPS generates audio and visual feedback directly related to the amount of external noise detected by the system. For example, a circle on the screen could change in diameter, as sound, particularly voice, grows louder. An "echo", like that heard in a stairwell, is an example of audio feedback. Distortions could be applied to change the perception of the sound returned to the subject. We explored visual, auditory and mixed (both visual and auditory) feedback due to cross-modal interference (Bonneh et al., 2008), commonly associated with ASD. Unlike existing software based communication treatments, discussed in the previous chapter, SIPS provides feedback and abstract representation of voice, rather than situating the visualization in a game with a concrete goal or objective.

We describe here the creation of SIPS, the metaphors of feedback, specific forms of feedback within each metaphor, and the implementation of the software package.

Some research, content, and text from this Chapter is reproduced from (Hailpern et al., 2009b; Hailpern, 2008)



(a) *Explosions*

(b) *Spinning Dots*

(c) *Line Circle*

(d) *Volume Circle*

Figure 6.1: *Examples of visualizations used in SIP*

6.1.1 Forms of Visual Feedback

SIPS visual feedback (Figure 6.1) consists of one of three possible types of graphical objects: circular/spherical, lines, or found images (e.g., picture of cartoon character). These objects can be presented in one of four types of motion metaphors; Falling, Spinning, Flashing and Stationary. Among the four categories, approximately 12 unique motion/pattern combinations were created; most can function with any type of object (circle, found image, etc).

6.1.1.1 Falling Visual Feedback

The falling metaphor simulates a gravitational pull on objects. This includes particle effects like water from a shower head or fireworks (Figure 6.1a). Consider a series of spheres falling, like rain, from the top of the screen to the bottom. If each sphere's diameter represented the volume of a sound, a subject could gain a temporal understanding of what sound they have made, and how their current sound differs. The falling metaphor was selected to leverage stimuli that garner interest from children with ASD (Baggs, 2007; Grandin, 2006; Mukhopadhyay, 2000).

6.1.1.2 Spinning Visual Feedback

The spinning metaphor moves objects in a circular or spiral pattern, outward, from the center of the screen (Figure 6.1b). Consider a series of circles (with a diameter based on the average volume in the past 2 seconds), which travel outward on a spiral path. The current sound remains in the center, and the subject can examine how their current sounds directly impact the circles which travel in a spiral. The spinning metaphor was selected to leverage stimuli that garner interest from children with ASD (Baggs, 2007; Grandin, 2006; Mukhopadhyay, 2000).

6.1.1.3 Flashing Visual Feedback

The flashing metaphor focuses on objects appearing and disappearing, giving a flashing or flickering effect (Figure 6.1c). Consider color circles appear in random locations on the screen, and the size of the color relates to the volume of the screen. This feedback would give an immediate sense of how loud a vocalization was, by the amount of color quickly appearing on the monitor. Flashing feedback was investigated due to its high energy, which often appeals to neurologically typical children.

6.1.1.4 Stationary Visual Feedback

The stationary metaphor (Figure 6.1d) contained a singular object, at the center of the screen, and changing in one respect based on the sound of a subject. Stationary objects were explored to focus on change in an object (size, color, etc.) rather than object motion.

6.1.2 Forms of Audio Feedback

SIPS provided two categories of audio feedback based on the sound produced.

6.1.2.1 1-to-1 Audio Feedback

1-to-1 feedback is sound produced by the interface was directly related to sound produced by the subject (e.g., echo, or pitch-shifted version of the subject's voice). Though there was a slight delay between source sound and feedback, but both input and output occur simultaneously. By modifying a subject's sound, and returning it back to them, the impact of different vocalizations can produce a variety of new sounds, which can encourage vocalization.

6.1.2.2 Reward Based Audio Feedback

Computer sounds were produced upon completion of subject's sound. Duration of reward sound was related to duration of sound produced (longer sound made by subject resulted in longer reward). Sound could be music or found-audio (e.g., from movie or TV show). By treating audio as a form of reward, subjects were encouraged to produce sounds of a variety of length, in order to create audio that they enjoyed hearing.

6.1.3 Implementation

SIPS was built using Java 1.4 and the Processing Visualization (Fry and Reas, 2007) toolkit. During testing and experimentation, SIPS was run on an iMac computer, with the Phoenix Audio SOLO microphone (noise canceling).

6.2 Research Questions

We pose the following research questions about the effects of contingent audio and/or visual feedback on low functioning children with ASD. Q1: Will at least one form of real time computer-generated feedback positively impact the frequency of spontaneous speech-like vocalizations? R1 is the primary question of SIP: testing the impact of computer-generated feedback. R1 builds upon the success of low-tech alternatives (e.g., image cards (Bondy and Frost, 2001), mirrors (Marshalla, 2001)) and other related work. The remaining research questions examine modes of feedback, and their implications on frequency of spontaneous speech-like vocalization. Q2-Q5 are derived from research into cognitive profiles of children with ASD (Leonard, 2000; Paul et al., 2007) concluding that individuals with ASD prefer visual feedback (Baggs, 2007; Grandin, 2006; Minshew et al., 1997; Mukhopadhyay, 2000). The responses to Q2-Q5 will directly impact future systems and the extent to which individualization is needed. Q2: Will all forms of feedback positively impact the frequency of spontaneous speech-like vocalizations? Q3: Will subjects increase the frequency of their spontaneous speech-like vocalizations in all conditions with visual only feedback, audio only feedback and/or mixed feedback? Q3a: If there is a modality that approaching or is significant (Q3), is there a specific form of that feedback in that modality that positively impacts frequency of spontaneous speech-like vocalizations? The quantitatively driven investigation of Q3 may hide the impact of a specific form of feedback. If that one form of feedback fails to significantly adjust the results in Q3, it will never be analyzed in Q3a.

Therefore; Q4: By testing feedback conditions that were qualitatively favored by subjects (assessed during experiment and via video), will we uncover forms of feedback that positively impact the frequency of spontaneous speech-like vocalizations? Q5: Is there a modality of feedback whose variations indicate (Q3, Q3a, and Q4) the child's frequency of spontaneous speech-like vocalization are positively impacted.

6.3 Within-Subject Experimental Design

Our subjects demonstrated limited response to requests or instructions to perform tasks due to the severity of their ASD. Therefore, engaging subjects in the same activity across trials and sessions was not a viable option. We relied on the visual/auditory feedback to be sufficiently engaging to promote spontaneous speech-like vocalizations. The feedback presented and tested was varied across children to enable an exploration of R3 and R3a. As a result, each child's performance served as his or her own baseline for comparison. Given the number of subjects participating and the questions generated, a within-subject design was selected. The analyses were conducted using a baseline created by each child and comparing that baseline to each of the computerized feedback conditions: visual, auditory or visual/auditory combined.

The within-subject experimental design (Kazdin, 1982) , an adaptation of the alternating treatments design (Barlow and Hayes, 1979), consisted of five non-verbal children (aged 3-8 years) diagnosed with "low-functioning" ASD. Each child enrolled in the study first participated in one to three 30-minute "orientation sessions" which acclimated the child to the study room, researchers, and computer feedback. No data were recorded during these sessions, though initial preferences for feedback type/style were noted.

Each child attended 6 data sessions after completing the orientation period. A data session lasted for approximately 40 minutes and consisted of approximately 8 two-minute trials. During a trial, a researcher exposed the subject to different forms of feedback (permutations of audio and visual). Each trial began with an antecedent demonstration by the researcher (e.g., saying "boo" and pointing to screen). The subject then could engage the system in whatever manner they chose.

Feedback permutations were selected based on qualitative vocalization frequency. Order of presentation was randomized across sessions to accommodate for order effects. However, the first trial of each session was a baseline trial with no audio or visual feedback. Although this baseline trial provided a means of comparison for assessing changes in spontaneous speech-like vocalizations due

(a) *projector with open room* **(b)** *projector with separated area* **(c)** *large screen computer at desk*

Figure 6.2: *Room Setup*

to visual/auditory feedback, we provided no control for order effects related to the presentation of the baseline condition.

6.3.1 Room Setup

Due to the varying personalities and ability to attend in a chair for an extended time, a variety of room configurations were employed. The room configuration was selected based on each child's preference and ability to sit in a chair, assed during the orientation sessions. Figure 6.2 illustrates the room configurations.

6.3.2 Independent Variables

Our within-subject experiment analyzed the dependent variable Spontaneous Speech-Like Vocalization (SSLV). A more detailed explanation of dependent variable selection is conducted in Chapter 5: A^3 Coding Guideline. The independent variables were the various permutations of visual and auditory

feedback. This facilitated contrast between the mode of feedback (visual, auditory, and mixed) as well as the different types of feedback (12 visual and 5 auditory forms).

Chapter 7

Spontaneous Speech-Like Vocalizations: Analysis, Results, and Discussion

Over a six-month period, 1200 minutes of video were annotated with all variables in the A^3 guidelines. However, we focused our analysis on Spontaneous Speech-Like Vocalizations (SSLV), one of the dependent variables from A^3. There is clear and important distinction between those vocalizations that are spontaneous and those that are imitative. This is critical when assessing children with special needs (Halle, 1987).

> *Spontaneous Speech-Like Vocalizations (SSLV)– sounds produced by the subject that could be phonetically transcribed (sounds that could be useful in oral communications) and are not being imitated.*

Unlike imitated vocalizations (echolalia), SSLVs are more indicative of vocalizations that may be used for meaningful speech because they rely on longer-term storage and retrieval of linguistic information.

7.1 Questions Analysis Methods

7.1.1 Dependent and Independent Variables

Our within-subject experiment analyzed the dependent variable Spontaneous Speech-Like Vocalization (SSLV). The independent variables were the various permutations of visual and auditory feedback. This facilitated contrast between the mode of feedback (visual, auditory, and mixed) as well as the different types of feedback (12 visual and 5 auditory forms).

Some research, content, and text from this Chapter is reproduced from (Hailpern et al., 2009b; Hailpern, 2008)

7.1.2 Question Analysis

Each subject was analyzed separately. Due to the varying lengths of each trial, a comparison between the number of occurrences of SSLV would be weighted towards longer sessions. To mitigate this effect, we analyzed a normalized frequency of SSLV (occurrences in trial divided by trial duration). Wilcoxon rank-sum and Kruskal-Wallis tests were used to compare the number of SSLV in response to different types of feedback. The Wilcoxon rank-sum test is a non-parametric alternative to the paired T-test. The Kruskal-Wallis test is a non-parametric alternative to a one-way analysis of variance (ANOVA). These tests were well suited for these data where distributions were not normal and where numbers were small because they do not make any distributional assumptions. All tests used a two-tailed alpha with a $p<0.05$ denoting statistical significance.

Q1 Analysis examines if there is at least one form of computer generated feedback that will positively impact a subject's frequency of SSLV. If there at least one condition in Q2-Q4 shows feedback has a positive impact on frequency of SSLV, we can conclude R1 is true for that subject.

Q2 Analysis indicates, in general, that all forms of feedback (regardless of mode/style) increase frequency of SSLV. Analysis of Q2 for each subject is determined by comparing the frequency of SSLV at baseline to frequency across all types of feedback using the Wilcoxon rank-sum test.

Q3 Analysis indicates if all forms of feedback in a specific modality positively impact SSLV. Analysis of Q3 for each subject is determined by performing a Wilcoxon rank-sum test comparing frequency of SSLV at baseline with frequency of SSLV in groups audio only, video only, and mixed feedback. Results from Q3 can be Video (video only significant $p<0.05$), Audio (audio only $p<0.05$), Mixed (mixed feedback only $p<0.05$) or some permutation of the three. If none have a significant p value, Q3 is considered Neither, indicating that no modality increased the frequency of SSLV (all $p>=0.05$).

Q3a examines if there is a specific type of feedback that increased frequency of SSLV in a modality that approached significance. Using the result from Q3, we will tease out specific forms/combinations of feedback within those statistically significant modalities (visual, auditory, mixed). Trials within the specific modality are broken down into subcategories based specific forms of feedback and tested against baseline using the Wilcoxon rank-sum test. Q3a is only asked if p values for Q3 were approaching statistical significance.

Q4 Analysis was performed using qualitative observations from researchers and video to further guide analysis. This enabled us to utilized overlooked forms of feedback that increased frequency

	Age	Diagnosis	Room Setup	Any Feedback	Visual Only	Audio Only	Mixed Feedback
Oliver	5	ASD	C	0.065 [-1.85]	0.386[-0.87]	0.063[-1.86]	0.058[-1.89]
Frank	8	ASD + Downs	C	0.024 [-2.26]	0.556[0.59]	0.011[-2.56]	0.006 [-2.71]
Larry	4	ASD + Downs	C	0.850 [-0.19]	0.796 [0.26]	0.805 [0.25]	0.650[-0.45]
Diana*	4	ASD	B	0.789 [-0.27]	0.016 [-2.41]	not used	0.470 [0.72]
Brian	3	ASD	A	0.834 [0.21]	0.766 [0.30]	not used	0.796 [-0.26]

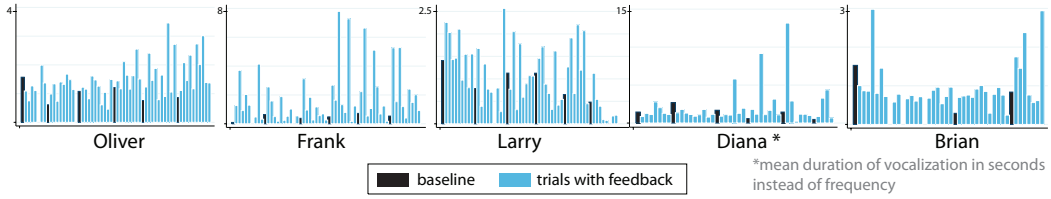


Figure 7.1: *Demographics and Frequency of SSLV*
Wilcoxon Rank-Sum Test from R2 and R3 Analysis
High level graphical comparison of Frequency of SSLV per 10 seconds across all trials for all subjects.

of SSLV. Using the Wilcoxon rank-sum test, we compared baseline with conditions that were qualitatively observed to increase SSLV frequency.

If significance was not found, the Kruskall-Wallis test was used to determine if differences existed in SSLV across feedback type, while excluding baseline measures. This additional analysis allows us to compare the impact of one form of feedback against all others.

Q5 Analysis required a categorization of the forms of feedback which illicit an increase in SSLV frequency. We extracted the mode of feedback found to have the most impact in Q3, Q3a and Q4. This synthesis of results provides a better understanding of what modes of feedback are engaging.

Using these questions as guides, we analyzed the results from each of our subjects to assess the impact of computer based visual and auditor feedback on SSLV. The following chapter describes the analysis in detail by presenting the findings broken up by subject.

7.2 Results

To protect the privacy of our subjects, we have changed their names; Gender status was maintained. All five of the subjects' spoken language developmental benchmarks (Paul, 2008) were in the first phase (Preverbal Communication),roughly equating to the development of a neurologically typical 6-12 month old.

	Found Audio	Echo
Audio Without Visual	0.200 [-1.28]	0.045 [-2.00]
Audio With Visual	0.082 [-1.74]	0.073 [-1.79]
Any Condition	0.076 [-1.77]	0.04 2[-2.03]

Table 7.1: *Comparison of Oliver's audio feedback*

7.2.1 Subject: Oliver

Initial analysis of Oliver's data (Figure 7.1) demonstrated borderline significance comparing baseline to all feedback (Q2). Further, the audio only and mixed feedback conditions (R4) approach significance. Due to a trend towards significance in the two conditions involving audio, we compared frequency of SSLVs at baseline with any condition containing audio feedback (both with and without visual feedback). There was a statistically significant difference between conditions containing any audio feedback and those containing no audio (p=0.045 [-2.00]). We conclude that audio feedback may have played a role in increasing the frequency of Oliver's SSLVs (Q5).

Since audio appeared to increase the frequency of Oliver's SSLVs, we explored impact of different forms of audio feedback in combination with visual feedback. Table 7.1 shows that echo feedback encouraged SSLV, while visual feedback did not appear to have significant impact on SSLV frequency (Q3a). We qualitatively observed that Oliver reacted positively to audio from a popular cartoon show. Our data confirms this by approaching statistical significance (p=0.083 [-1.74]) (Q4).

From this analysis, we conclude that Oliver increased his frequency of SSLV in conditions with audio feedback. Specifically, he increased SSLV in conditions with echoing audio feedback (Q1).

7.2.2 Subject: Frank

Initial analysis of Frank's data (Figure 7.1) showed a significant difference in frequency of baseline SSLVs and frequency of SSLVs with all feedback (Q2). We found a statistically significant difference in frequency of SSLVs with audio only and mixed feedback (Q3). Due to significance in both conditions

Audio Feedback	p value with visual feedback	p value without visual feedback
Any Found Audio	0.010 [-2.57]	0.011 [-2.56]
Child's Cartoon Found Audio	0.003 [-2.98]	0.011 [-2.56]]
Echo	0.005 [-2.80]	No data

Table 7.2: *Comparison of Frank's audio feedback*

Form of Visual Feedback in Addition to Audio	P Value
No Visual Feedback	0.011 [-2.56]
Cartoon Image	0.046 [-2.00]
Firework-like	0.004 [-2.86]
Spinning Spiral of Dots	0.160 [-1.41]
Fast Flash	0.010 [-2.58]
Line Circle	0.032 [-2.14]
Random Dots	0.134 [-1.50]
Shower	0.046 [-2.00]

Table 7.3: *Frank: Form of visual feedback with any audio*

with audio, we compared frequency of baseline SSLVs with any condition with audio feedback. There was a highly significant association between audio feedback and SSLVs (p = 0.004 [-2.84]) (Q5).

Given the robust effect of audio feedback, we compared Frank's responsiveness to audio feedback with and without visual feedback (Table 7.2). Audio feedback was categorized as "found audio" and "echo". Based on our qualitative observations, we isolated and analyzed trials where audio feedback from a specific child's cartoon was present. Frank demonstrated the most significant increase in frequency of SSLVs over baseline when audio from the cartoon was present (Q3a, Q4). For this subject, visual feedback had a positive impact on the frequency of SSLVs when audio was also present.

Finally, we examined all conditions with audio feedback into specific forms of visual feedback to assess the impact of different forms of visual feedback on the frequency of SSLV production. Based on qualitative observations, we analyzed trials where a visual image from a specific cartoon was present. Frank demonstrated increased SSLV frequency over baseline for all visual feedback in addition to audio for all but Spinning Spiral of Dots and Random Dots (Table 7.3), with the highest significance in Firework-Like Feedback (Q3a).

From this analysis, we conclude that Frank had a higher frequency of SSLV to conditions with audio feedback and both audio and visual feedback together (Q1 , Q5). Specifically, he appeared to show increased SSLV when audio and visuals from a specific cartoon. Interestingly, his mother stated that Frank did not watch this cartoon show.

7.2.3 Subject: Larry

Initial analysis of Larry's data (Figure 7.1) failed to reach statistical significance (Q2, Q3). While formal statistical tests did not reach statistical significance, qualitative observations from researchers and

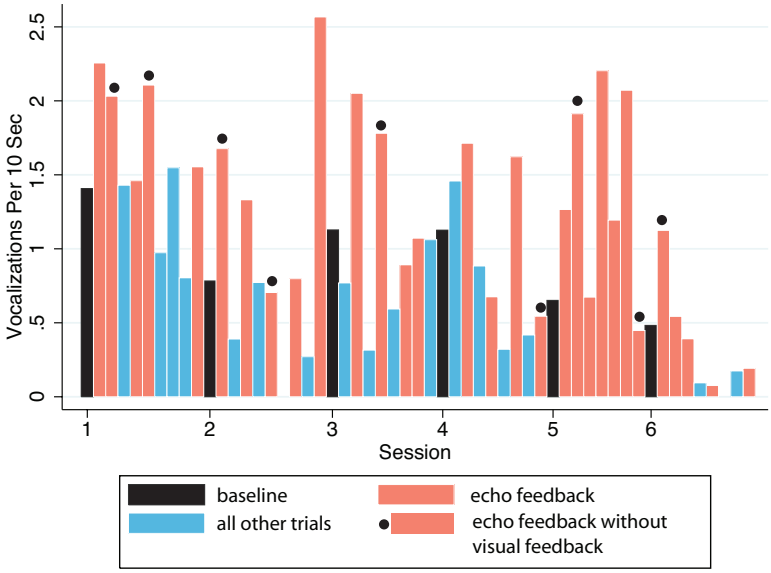


Figure 7.2: *Larry's SSLV frequency, by session and trial*

study video, in conjunction with graphical representation of the data (Figure 7.2) led us to believe that there was feedback that had impact on frequency of SSLV, specifically conditions with echoing audio feedback. Qualitatively, researchers observed a higher degree of attention and SSLV, during conditions with echo/reverb feedback.

Comparing conditions with echoing feedback with baseline produced a lower p-value than other analysis (p=0.243[-1.16]), yet it did not reach p<0.05. To examine the impact of echoing feedback, we repeated our analysis across test conditions. We performed a Wilcoxon rank-sum test to compare conditions using echoing feedback with visual feedback to conditions with only echoing feedback and no visual feedback. Given p=0.970, we concluded that there was no significant difference in SSLV between echoing conditions with and without visual feedback.

	Baseline	**Any Condition with Echo**	**Audio Only**	**Mixed + Visual Only**
Baseline	—	—	—	—
Any Condition with Echo	0.284 [-1.07]	—	—	—
Audio Only	0.055 [1.91]	0.034 [2.13]	—	—
Mixed + Visual Only	0.410 [0.83]	0.023 [2.27]	0.396 [-0.85]	—

Table 7.4: *Larry's comparative conditions (Row vs. Col)*

Form of Visual Feedback in Addition to Audio	P Value (without audio)	P Value (with audio)
Firework-like	0.136 [-1.49]	0.934 [-0.08]
Spinning Image	0.020 [-2.32]	0.201 [-1.28]
Shower-like	—	0.439 [0.78]
Fast Flash	—	0.739 [-0.33]
Multiple Circles	0.020 [-2.32]	0.556 [-0.59]
Line Circle	0.617 [0.50]	0.439 [0.78]
Fast Spin	—	0.617 [0.50]
Found Imagery	0.003 [-2.97]	0.330 [-0.97]

Table 7.5: *Diana: Forms of Visual Feedback tested, vs. baseline (with and without audio)*

To compare the impact of echoing feedback on SSLV with other forms of feedback, we used the Kruskal-Wallis test. First, we categorized all of Larry's trials into one of the following 5 conditions; (1) baseline, (2) any condition with echoing feedback, (3) only audio feedback (excluding echoing), (4) only visual feedback (excluding echoing), (5) audio + visual feedback (excluding echoing). The Kruskal-Wallis test had a p=0.060. To increase statistical power, we collapsed groups[1] by combining visual only feedback with mixed condition since groups had visual presentations (comparative analysis between collapsed groups: Wilcoxon rank-sum p=1.000[0.00]). Analysis of these groups found a statistically significant difference (p=0.030 by Kruskal-Wallis test). A post hoc pair-wise comparison of each condition, using Wilcoxon rank-sum test (Table 7.4) was performed. Statistically significant differences were found between the echo condition and audio only(visual + mixed) (p=0.034, p=0.023 respectively) (Q4).

From this analysis, we conclude that Larry showed preference for echoing audio feedback (Q1, Q5). However, we believe that with more statistical power, we could make a more conclusive statement.

7.2.4 Subject: Diana

Diana responded to many commands by her mother such as sit, stop, come here, and wait. Diana demonstrated two to three signs for communication (e.g., more, music), though articulation of signs was poor, and frequency was low (about 1 per session).

Initial data analysis for Diana, found much higher p values (0.5-0.9) than expected when comparing them to qualitative notes made by researchers. Confused by these findings, we examined annotations made by video coders and noticed that large strings of Diana's SSLVs were being grouped together.

[1] Collapsing two groups increases the number of data points in the resulting group, thus increasing the statistical power during comparison.

A^3 guidelines stated that utterances must be separated by a pause of 2-seconds to be considered independent. However, Diana's pauses ranged from 1-to-1.5 seconds in duration. As a result, phrases of multiple utterances were captured as just one occurrence. To accommodate her shorter pauses, we re-analyzed her data using mean duration of SSLVs rather than frequency. For this subject, we used average duration as a proxy for frequency.

Initial analysis of duration of SSLV (Figure 7.1) showed significance for visual only conditions (Q2, Q3). Audio only feedback was not used, due to lack of interest observed in initial orientation sessions.

To examine impact of visual feedback, we broke down the forms of visual only feedback and compared average duration of spontaneous SSLVs with those produced in baseline condition (Table 7.5). The last row in Table 7.5 is an amalgam of different forms of visual feedback in which abstract colored dots are replaced with one or more found image(s). This data support our qualitative observations that Diana only responded to conditions where images shown were from cartoon shows, and that audio feedback reduced her SSLV (Q4). Three statistically significant conditions were Spinning Image (a found image from a cartoon spins on axis), Multiple Circles (many dots or found images appear on screen; size based on volume of sound produced) and any feedback with Found Images (there are overlaps between groups) (Q3a, Q4).

From this analysis, we conclude that Diana produced more SSLVs (mean duration) with visual feedback compared to baseline and mixed (Q1, Q5). Specifically, she appeared to show increased engagement with forms of visual feedback that contained a cartoon character (though a specific preference did not appear). Diana was reported to watch movies/TV-shows with these characters.

7.2.5 Subject: Brian

Brian was the most difficult subject for us to qualitatively discern a particular pattern or "taste" for feedback. This was supported by extremely high p-values for all coarse tests conducted on the other subjects (Figure 7.1). During three sessions, we inadvertently failed to run a baseline, reducing the number of comparison points to three instead of six. This reduced statistical power. While Wilcoxon rank-sum statistics approached significance for one particular form of visualization in which a cartoon character spun in a circle centered on screen, it failed to reach significance.

From this analysis, we could not conclude that Brian had a significant reaction to any form of feedback (either compared to baseline or against each other) (Q1-Q5).

	R1	R2	R3	R4	R5
Oliver	▲	▼	▼	▲	**A**
Frank	▲	▲	**A + M**	▲	**A + M**
Larry	▲	▼	▼	▲	**A**
Diana	▲	▼	**V**	▲	**V**
Brian	▼	▼	▼	▼	▼

▲ *Positive* ▼ *Negative*

A *Audio* **V** *Visual* **M** *Mixed*

Table 7.6: *Results by subject*

7.3 Discussion

After a thorough examination of the quantitative data collected, we are able to summarize the findings in relation to our 5 questions (Table 7.6).

7.3.1 Q1 Discussion

In 4 of the 5 subjects, we found that at least one form of feedback created an increased frequency of SSLVs. We were unable to show that any form or modality of feedback, when compared to baseline, significantly increased the frequency of SSLVs for Larry and Brian. This may be, in part, due to the small number of data points collected and high degree of ASD. We were, however, able to demonstrate that echoing audio feedback produced a significant difference in frequency of SSLVs when compared with all other forms of feedback for Larry. Overall, we conclude that feedback may encourage SSLV in children with ASD.

7.3.2 Q2 Discussion

Only one of five subjects found all forms of feedback, regardless of mode or form, to have a positive impact on frequency of SSLV. This finding suggests that not all forms of computer feedback work for all children.

7.3.3 Q3 & Q5 Discussion

It is commonly believed that individuals with ASD respond better to visual feedback than auditory (Baggs, 2007; Grandin, 2006; Mukhopadhyay, 2000). However, we had two subjects who responded

primarily to auditory feedback (Oliver and Larry). One preferred a mixed condition (Frank). One responded to visual only (Diana). One subject (Brian) did not show any significant reaction to any form of feedback. When taken from a more global level, 3 of 5 subjects responded to audio feedback, and 2 of 5 responded to visual feedback Table 7.6. This suggests that further exploration of feedback in both visual and audio modality is essential. This finding is of particular note in that it is in contrast to other work.

7.3.4 Q3a & Q4 Discussion

Though some subjects had a larger range of forms of feedback that resulted in increased frequency of SSLV than others, 4 of 5 subjects did have one particular condition that out-performed the others. The specific results, in conjunction with varied modes of feedback that resulted from Q3 analysis, indicate that visualizations, and any potential therapeutic application, will likely need to be tailored to individual subjects. The degree of customization is unknown due to small sample size. We can proceed, however, knowing that individual interests/preferences must be taken into consideration. This work illustrates the varied forms of audio/visual feedback that garnered the increase in SSLV.

7.3.5 Parental Response

In addition to data from subjects during the sessions, we asked for anonymous parental response in the form of a written questionnaire. Feedback from parents was positive and encouraging. Parents responded with high praise for our technique, and asked for similar solutions to be put to use in their own homes. One mother stated,

> *My child's reaction is one of excitement and looking forward to see what was next to come. Applause on your study. You may be onto something here.*

Another mother stated her child's reaction,

> *Since my son is fairly severely affected by autism, he stays in his "own little world" quite a bit. So the fact that he showed interest in and seemed to enjoy some of the visuals and sounds is quite a positive thing. Thank you.*

7.4 Summary of Findings

Given the results from the SIP study, we believe that audio and/or visual feedback can be used to encourage spontaneous speech-like vocalizations in low-functioning children with ASD. In addition, SIP demonstrated that both visual and auditory feedback can impact spontaneous speech-like vocalization. This suggests that further exploration of feedback in both modalities is essential. This finding is of particular note in that it is in contrast to other existing work.

SIP also suggests that low-functioning children with ASD may have distinct and varied preferences for types/styles of feedback. As a result, individual customization may be necessary in future efforts. Although the range of variation necessary is unknown, the final solution might include a suite or menu of feedback styles that may be selected by the parent, clinician, or child.

Chapter 8

Speech Like & Non-Speech Like Vocalizations: Data Analysis, Results, and Discussion

This chapter seeks to examine the influence of and relationship between Computer Feedback Systems (CFS) on both speech-like vocalizations (sounds that can be phonetically transcribed) and non-speech-like vocalizations (e.g. grunt, screech, etc). Ideally, for purposes of speech development, we suggest that the goal should be to increase speech-like vocalizations in the presence of CFS with a decrease in or null effect on non-speech-like vocalizations. It should be noted that SIP was not explicitly designed to impact non-speech-like vocalizations (positively or negatively).

By definition, the categories of speech-like and non-speech-like vocalization are mutually exclusive: children can produce one form of vocalization at a time. Consequently, one might anticipate that an increase in one type of vocalization would lead to a decrease in the other. However, within a specified time period, it would be possible to increase the frequency of both types of vocalizations if the initial baseline rate of vocalization is relatively low. It is, therefore, feasible that the CFS may increase the frequency of both speech-like and non-speech-like vocalizations, especially since both types of vocalizations elicited visual and/or auditory feedback from the system.

The main contribution of the present work is to assess the differential impact of CFS on speech-like and non-speech-like vocalizations as well as their interaction and impact on each other. Such information is critical in evaluating the effectiveness of CFS and guiding the development of future feedback systems.

8.1 Questions & Analysis Methods

8.1.1 Dependent and Independent Variables

Our analyses are organized around four specific questions that tease apart the relationship between CFS, speech-like vocalizations, and non-speech-like vocalizations. Dependent measurements are centered

Some research, content, and text from this Chapter is reproduced from (Hailpern et al., 2009a)

on Speech-Like Vocalizations (SLV) and Non-speech-Like Vocalizations (NSLV). Chapter 5 provides further details on and definitions of SLV and NSLV. *Note that a subset of SLV (those vocalizations that were not imitative) was the only dependent variable examined in the previous chapter.*

Trials included varying forms of visual and auditory CFS. Subsequent to baseline trials, each trial included at least one form of visual or auditory feedback. Some trials included both auditory and visual feedback (Mixed).

8.1.2 Questions

8.1.2.1 Q1: Does CFS Impact Children's Frequency and/or Duration of Vocalization?

Question 1 examines, at the most general level, the impact of CFS on vocalizations in low functioning children with ASD. To conduct this analysis we compared baseline trials (no feedback) to all other trials (regardless of feedback style or modality). We assessed the impact on three dependent variables: rate of SLV, average duration of SLV, and rate of NSLV. Because our data does not contain NSLV duration, we were unable to assess the impact on NSLV duration. Upon further consideration, this may have been a difficult variable to assess in that many NSLVs are nearly instantaneous (e.g. lip pop), and thus difficult or impractical to measure.

8.1.2.1.1 Desired Outcome Our desired outcome is that CFS would increase rate and/or duration of SLVs while decreasing the rate of NSLVs. In other words, we hope that CFS would encourage children to vocalize in a manner that could be used for speech, while discouraging non-speech vocalizations (e.g. screaming, grunts, etc) that may be viewed as less socially acceptable. This is based on how CFS was introduced to each child; specifically the CFS was demonstrated for each child through use of speech. Specifically, the investigators would draw each child's attention to the screen while saying words such as the child's name or an exclamation (e.g., "Boo!"). Although imitation is an area of difficulty for many children with ASD (Jones and Prior, 1985), this does not mean the skill is nonexistent. Imitation is a commonly employed strategy used by children and adults in unfamiliar situations.

8.1.2.2 Q2: Does CFS Differentially Increase SLV Relative to NSLV?

Our second question examined potential trade-offs between rate of SLV, and rate of NSLV. By examining the overall change in vocalizations from both sides of the vocalization spectrum, we can best examine

the overall impact of CFS. This analysis will follow Q1 by comparing baseline to all trials, while examining the effect of CFS through a different dependent variable, the ratio of SLV to NSLV:

> *SLV:NSLV Ratio – This dependent variable measures the relative change between potentially phonetically transcribable vocalizations and those that are not.*

An increase in one form of vocalization may lead to a decrease in another form or both types may increase as an expression of decreased silence.

8.1.2.2.1 Desired Outcome Consistent with the rationale we presented in relation to Q1, our desired outcome is less frequent NSLV compared to SLVs, indicating an increase of SLV compared to NSLV during CFS trials. Thus our ratio should get larger in non-baseline conditions.

8.1.2.3 Q3: Are the different vocalization measures correlated?

To better understand if there is a significant relationship between SLV rate, SLV duration and/or NSLV rate, we conducted correlations between the different dependent variables. We examined the relationship between dependent variables within each child.

8.1.2.3.1 Desired Outcome We hope that there will be a negative correlation between speech-like vocalizations and non-speech-like vocalizations indicating that in the presence of one, the other decreases.

8.1.2.4 Q4: How does Modality Impact Vocalizations

To examine the differences between different modalities of feedback (audio/visual), we sought to re-examine Q1 and Q2 while comparing baseline to conditions with Audio Only (no visual feedback), Visual Only (no audio), and Mixed (both visual and auditory feedback must be present).

- Does the impact of CFS on children's vocalizations (SLV and NSLV) differ as a function of feedback modality (audio, visual, mixed)?
- Does the impact of CFS on children's SLV:NSLV ratio differ as a function of feedback modality (audio, visual, mixed)?

	SLV-R (%)	NSLV-R (%)	SLV-D (%)
Maximum Power	100	99	96
Minimum Power	5	6	9
Mean Power	35	41	34
Median Power	12	35	26

Table 8.1: *Statistical Power By Measure*

8.1.2.4.1 Desired Outcome We hope that the impact of CFS would differ as a function of modality. Though the common thought is that children with ASD respond more favorably to visual stimuli, based on the previous findings of Hailpern et al., we anticipated both audio and visual feedback would produce effects that would differ for individual childre

8.1.3 Statistical Tests

Given the non-parametric nature of the data collected, traditional paired t-tests would not be applicable. The Wilcoxon Rank-Sum was used as a non-parametric alternative. Similarly, the Spearman rank correlation test was used as a non-parametric alternative to the Pearson correlation test.

8.1.4 Significance

All tests used a two-tailed alpha with a $p<.05$ to denote statistical significance. We indicate statistical significance with an asterisk ($*$). While performing multiple comparisons may suggest statistical adjustment to a more conservative value (i.e., Bonferroni correction), the small number of data points and low statistical power of our study would suggest a more relaxed threshold (See Table 8.1). Taking both perspectives into consideration, an alpha of $<.05$ was determined to be a reasonable approach. Moreover, as a preliminary pilot study, the focus of this research was to highlight avenues of future research rather than to test competing hypotheses. Consequently, following the rationale of Savitz (Savitz and Olshan, 1995), we believe that the potential of detecting false positives (5%) was outweighed by the concern of missing meaningful effects. For all results and discussion, a $p<.05$ will be used to denote significance.

Nonetheless, we performed a sensitivity analysis in which p-values were adjusted using a Bonferroni correction for multiple rank-sum comparisons. For the dependent variable SLV-D, we performed two statistical comparisons on each independent data set, resulting in an adjusted threshold of .0250 for statistical significance. Similarly, the SLV-R and NSLV-R variables were used in four statistical

comparisons on each independent data set, resulting in a threshold of .0125 for statistical significance. We indicate statistical significance for the sensitivity analysis with a dagger (†) in Table 8.2 and Table 8.4.

8.2 Results

We present here the results related to each research question. We present a summary of the results in addition to the statistical findings. Due to the within-subject design of this experiment, we will further present the results for each child independently. A detailed discussion of the results and their implications is presented in the following section.

8.2.1 Q1: Does CFS Impact Children's Rate and/or Duration of Vocalization?

> *CFS appeared to influence the rate or duration of vocalizations in four of the five children. (See Table 8.2)*

With all feedback types collapsed, Oliver significantly decreased NSLV-R while our analysis trended towards increasing his SLV-R with p=0.055. Frank significantly increased both SLV-R and NSLV-R in conditions with CFS. Larry significantly decreased his SLV-D during CFS. Diana showed no change in NSLV-R, SLV-R, or SLV-D. Brian significantly decreased NSLV-R with CFS. Consequently, when all feedback trials were combined desirable effects of either increasing SLV or decreasing NSLV were observed in 3 out of 5 children with only one child showing an undesirable decrease in SLV-D.

8.2.2 Q2: Does CFS Differentially Increase SLV Relative to NSLV?

> *When the overall impact of the CFS is considered, only two out of five children increased their ratio of SLV-R relative to NSLV-R, but primarily by decreasing the freq of NSLV. (See Table 8.2)*

Oliver significantly increased SLV-R:NSLV-R ratio through a decrease of NSLV-R. Frank had no significant change when comparing SLV-R to NSLV-R. It is of interest, however, that Frank increased both his SLV-R and NSLV-R, resulting in a net ratio gain. Larry and Diana presented no significant change in relative use of SLV-R to NSLV-R, at least when results from all feedback types were collapsed. Brian significantly increased his SLV-R per NSLV-R primarily through a decrease in NSLV-R during

	NSLV-R			SLV-R			SLV-D			SLV-R/NSLV-R		
Oliver	2.24,	0.025*	▼	-1.92,	0.055		1.17,	0.240		-2.82,	0.005*†	▲
Frank	-2.38,	0.017*	▲	-2.51,	0.012*†	▲	1.07,	0.286		-1.66,	0.098	
Larry	0.62,	0.535		-0.22,	0.829		1.99,	0.046*	▼	-0.18,	0.861	
Diana	0.99,	0.324		-0.10,	0.920		-0.27,	0.789		-1.32,	0.188	
Brian	2.08,	0.038*	▼	0.12,	0.907		1.54,	0.124		-2.03,	0.042*	▲

Table 8.2: *Wilcoxon Rank-Sum across all trials.*
Rate (-R) & Duration (-D) of SLV and NSLV. (z value, p value).
Arrow represents statistically significant change/direction in measure/condition vs. baseline
** denotes significance < .05*
† denotes significance in Sensitivity Analysis (Section 8.1.4)

	SLV-D by SLV-R		SLV-R by NSLV-R		NSLV-R SLV-D	
Oliver	-.451,	<.001*	-.049,	.710	-.198,	.138
Frank	-.430,	.001*	-.399,	.002*	.027,	.846
Larry	.278,	.048*	.208,	.135	.099,	.488
Diana	.179,	.249	-.168,	.282	.208,	.182
Brian	.125,	.429	.334,	.027*	-.122,	.440

Table 8.3: *Spearman Rank Correlations (rho, p)*
rho indicates relative slope between variables, and effect size
** denotes significance < .05*

CFS. Note that findings from the omnibus test may mask important individual differences in children's response to particular modalities/forms/styles of feedback as shown in Chapter 7.

8.2.3 Q3: Are the different vocalization measures correlated?

When considering the correlations among the three dependent variables, we noted significant relationships between SLV-D and SLV-R for three of the five subjects and significant relationships between SLV-R and NSLV-R for two out of five subjects. (See Table 8.3)

Oliver and Frank show a negative correlation between SLV-D and SLV-R, while Larry's correlation was positive. Further, both Frank and Brian's SLV-R and NSLV-R correlations were significant, however, the direction of the association was in opposite directions from each other.

8.2.4 Q4: How does Modality Impact Vocalizations

When considering the effect of feedback modality on vocalization, each child appeared to have a unique response to different forms of feedback. Specifically, two of the five subjects had a significant response to only audio feedback; three of the five subjects had a significant response to conditions with only visual feedback, and all five subjects responded significantly to the mixed feedback condition. The following subsections and Table 8.4 detail the responses of each child.

Consistent with the negative correlation in Table 8.3, Oliver significantly increased his SLV-R and decreased his SLV-D in conditions with only audio feedback. In addition, Oliver's SLV-R to NSLV-R ratio increased in both the auditory only and mixed conditions. This suggests that for Oliver audio

		NSLV-R			SLV-R			SLV-D			SLV-R:NSLV-R		
Visual Only	Oliver	1.52,	0.129		-0.87,	0.386		-0.82,	0.409		-1.68,	0.093	
	Frank	-2.95,	0.003∗†	▲	0.47,	0.637		-1.29,	0.196		2.01,	0.044∗	▼
	Larry	-1.29,	0.197		0.26,	0.796		0.33,	0.739		1.29,	0.197	
	Diana	-0.10,	0.920		0.90,	0.366		-2.41,	0.016∗†	▲	-0.40,	0.692	
	Brian	2.09,	0.037∗	▼	0.19,	0.850		1.34,	0.182		-2.04,	0.041∗	▲
Audio Only	Oliver	1.57,	0.116		-2.14,	0.032∗	▲	2.00,	0.046∗	▼	-2.72,	0.007∗†	▲
	Frank	-0.85,	0.394		-2.56,	0.011∗†	▲	0.00,	1.000		-2.57,	0.010∗†	▲
	Larry	1.32,	0.187		0.16,	0.869		1.57,	0.117		-0.54,	0.591	
	Diana§												
	Brian§												
Mixed Feedback	Oliver	-1.93,	0.054		-0.33,	0.742		1.40,	0.161		-2.74,	0.006∗†	▲
	Frank	-3.00,	0.003∗†	▲	1.70,	0.089		1.70,	0.089		-2.32,	0.021∗	▲
	Larry	-0.44,	0.660		1.03,	0.304		2.10,	0.036∗	▼	-0.20,	0.839	
	Diana	-0.54,	0.588		-2.14,	0.032∗	▲	0.72,	0.470		-1.64,	0.100	
	Brian	-0.26,	0.796		1.64,	0.101		2.24,	0.025∗†	▼	-1.51,	0.131	

Table 8.4: *Wilcoxon Rank-Sum across modalities.*
Rate (-R) & Duration (-D) of SLV and NSLV. (z value, p value)
arrow represents statistically significant change/direction in measure/condition vs. baseline
∗ denotes significance < .05, † denotes significance in Sensitivity Analysis (Section 8.1.4)
§audio only feedback was not tested due to lack of interest observed in initial orientation sections

	NSLV-R		SLV-R		SLV-D		SLV-R: NSLV-R	
	▲	▼	▲	▼	▲	▼	▲	▼
Visual Only	F	B	–	–	D	–	B	F
Audio Only	–	–	OF	–	–	O	OF	–
Mixed	F	–	D	–	–	LB	OF	–
# of Unique Subjects	**1**	**1**	**3**	**0**	**1**	**3**	**3**	**1**

Table 8.5: *Subjects With Significant Changes Compared to Baseline in Different Conditions*
First initial of subject's name represents change in behavior relative to baseline (arrows represent change direction)

feedback was a useful form of CFS. While the presence of visual feedback with the audio did not change his SLV-R or NSLV-R, it did improve the ratio between the two.

Frank significantly increased his NSLV-R in all conditions with visual feedback (visual only and mixed). The absence of change in NSLV with only audio suggests that visual feedback was the primary catalyst for increasing NSLV-R. Further, Frank increased his SLV-R in the conditions with only audio feedback. The absence of an increase in SLV-R in conditions with visual feedback suggests that Frank responded favorably to audio feedback and unfavorably to visual feedback in terms of desired responses. Additional support for this conclusion comes from the positive increase in Frank's SLV-R to NSLV-R ratio in conditions with audio feedback (audio only and mixed) paired with the decreased ratio in conditions with visual only feedback. Together these results suggest a cohesive picture for Frank with audio feedback serving as the primary catalyst for SLV, and visual feedback as the primary catalyst for NSLV.

Larry significantly decreased his SLV-D in the mixed feedback condition. Given the lack of significance found in visual only and audio only conditions, the data suggest that Larry's decrease in SLV-D is primarily due to the presence of audio and visual feedback together. Overall, Larry does not appear to be a good candidate for this particular CFS based therapy.

Diana significantly increased her SLV-D in conditions with only visual feedback, and significantly increased her SLV-R in the presence of mixed feedback. This indicates that Diana responded primarily to visual feedback, in particular a specific type of visual feedback (see Chapter 7 for details).

Brian decreased his NSLV-R in conditions with only visual feedback. This indicates that visual feedback may have suppressed his NSLV. Consistently, Brian's SLV-R:NSLV-R ratio increased during conditions with visual-only feedback. While not enough data were collected to determine the effect of audio-only feedback, the lack of a statistically significant response in all other conditions does suggest that audio feedback did not play a role in improving his ratio.

8.3 Discussion

indicates that for three of five of subjects adjusted the ratio changed in the presence of CFS. However, for the five conditions in which there was an improved ratio (more SLV-R per NSLV-R), one was due to suppressed NSLV-R and two were due to increased SLV-R. Our data suggest that CFS had a significant impact on all five subjects for at least one measure. However, the specific outcome varied by subject and by modality. This indicates that individual differences in subjects' response to computerized feedback should be considered in any attempt to develop feedback for the purpose of facilitating speech. In particular, the inclusion of audio as a significant modality is noteworthy here as the "conventional wisdom" is that visual feedback is particularly engaging to individuals with ASD (Minshew et al., 1997; Mohamed et al., 2006; Peterson et al., 1995). Although highly preliminary in nature, findings suggest that audio feedback should not be automatically discounted as a means to facilitate speech development, particularly if the audio feedback is shaped in a way that is not available in everyday encounters.

Table 8.5 illustrates the breakdown of the number of subjects whose vocalizations significantly changed with CFS. From this high level analysis, we are brought back to our basic questions regarding suppression vs. encouragement of SLV and NSLV, as well as the relationship between the two variables. Of particular note, no feedback conditions appeared to suppress SLV-R, which is critical to CFS being useful in a therapeutic condition.

Further, the data suggests that in certain conditions there may be a trade-off between SLV rate and SLV duration. When we examine the effect of CFS on SLV-D (particularly in Table 8.5), an effect of suppression appears, rather than encouragement. When considered together increases in SLV-R, evidence of this trade-off appears. This is further supported by examination of Table 8.3 and the negative correlations of Oliver and Frank (the exception being Larry with a positive correlation). In other words, there may be a trade-off between rate and duration of SLV: as subjects vocalize more, the individual vocalizations may become shorter. Consequently, thought should be given, both by investigators and by clinicians, as to the specific goals of their intervention, and recognize that some outcomes may in fact compete with one another. For researchers moving forward with similar work, this finding suggests that it may be naive to only examine rate or duration in isolation. Rather, an examination of both may be necessary so as not to miss the potentially important underlying effects of their CFS.

In addition to the potential for differentiated effects on vocalization rate and duration, recognizing the potential to differentially impact speech-like versus nonspeech-like vocalizations is key. Feedback

appeared to be effective in suppressing NSLV-R relative to SLV-R. This result suggests that the vocalizations elicited by CFS may be more adaptive for teaching spoken language skills. It is hard to fully understand the source of this relative change given the limited data of the study, which leads us to caution that the effects of any particular feedback on vocalization should not be assumed to be positive. The differential influence of visual versus audio feedback for Frank highlights this point, with the former facilitating nonspeech-like vocalizations and the latter speech-like vocalizations. That being said, CFS generally appeared to affect SLV-R more so than NSLV-R across the five children (see Table 8.5). Such results provide hope that CFS can encourage increased SLV-R vocalizations while decreasing or having minimal impact on NSLV-R.

Given our focus on statistical significance, we want to explicitly note that the absence of significance in a specific condition/modality does not indicate lack of potential impact. In Chapter 7, it was shown that specific feedback modalities and individual styles within that modality were often most successful for individual children. This finding was similarly noted in the qualitative analysis, particularly in regard to specific modalities. Yet Hailpern et al.'s study did not design or account for subject preference. In other words, the ideal form of feedback for each child in each modality may not have been explicitly built into the system. While our results do suggest conditions where CFS had impact, researchers should not rule out any of our conditions that did not have significant findings; the nature of this particular CFS may not have been engaging to that individual child.

8.4 Summary of Findings

Given the findings of this data analysis, we believe that Computer Feedback Systems (CFS) present an exciting and potentially promising new technique for encouraging vocalization in low functioning children with ASD. More specifically, CFS appears to encourage the rate of vocalizations that can be considered Speech-Like (sounds that can be phonetically described), while generally having little impact on those vocalizations that we call Non-speech-Like (e.g. grunt, screech, etc). In other words, CFS has the potential to differentially impact (favorably) speech-like versus non-speech-like vocalizations. However, the nature of this potential relationship needs to be directly considered rather than implicitly assumed for any particular child, form or modality of feedback. This understanding strengthens the potential impact of CFS, and the original study Chapter 7 & refchap:sip from which this data was collected. Together, these studies help delineate the need for future research in computerized feedback systems in order to better understand its impact on vocalization and potential in helping teach language and communication skills to children with Autistic Spectrum Disorder.

8.5 Follow-up Wizard-Of-Oz Study

One future application of the SIP framework, is to encourage specific forms of vocalization, namely teaching vocabulary acquisition and pronunciation. Given the current state of SIPS, this particular functionality is not readily available. However, a common technique used in design is a "Wizard-of-Oz" study. This is a form of experimentation in which a researcher operates a partially functional system to make it appear to be fully functional to subjects.

Researchers constructed a Wizard-of-Oz system, based on SIP, geared towards teaching specific skills. The model followed a common form of Behavioral Therapy (Lovaas, 1977): Prompt for word Ð wait for response Ð reward if correct or repeat if incorrect. We replaced the computer voice recognition with a researcher to test the concept. Researchers qualitatively noted Frank's response as being exceptional, both in terms of his reaction to the computer feedback and his eagerness to participate. Noting this, researchers asked his mother's permission to include him in an additional Wizard-of-Oz study.

This system aurally prompted subjects with a word in the form of the phrase "Say [word]." Once the prompt was completed, the computer provided visual feedback (spinning spiral of dots) and audio feedback (echo). Immediate feedback provided the subject with an instantaneous reaction to their sounds, for both visual and auditory reinterpretation. If the Frank did not repeat the sound, or the repeated sound was not "close enough," the researcher directed the system to re-prompt. If Frank's response was "close enough," the researcher directed the system to provide an auditory and visual reward.

With parental permission, we conducted 2 sessions using this system. The first consisted of 10 words, which had been previously used by Frank (according to his mother). Initially, Frank played with the system (similar to SIP sessions). After 15 minutes, he began repeating words upon the request of the system. At the end of the 30-minute period, Frank repeated every prompted word.

During the second session, we used 6 words his mother stated he had not spoken before, in addition to 4 words from the previous session. We asked Frank's mother to provide us with words she hoped he would learn, but has not used to date. Frank readily played the Prompt-Repeat game and attempted to repeat the new words. Though articulation was often unclear he made a concerted effort to repeat all 10 words, including the 6 new ones. Of particular note, Frank has been highly resistant in the past with this form of Vocal Imitation Language therapy.

Chapter 9

VocSyl: Syllable Project

The ability of children to combine syllables represents an important developmental milestone. This ability is often delayed or impaired in a variety of clinical groups including children with autism spectrum disorders (ASD) and speech delays (SPD). Prior work has demonstrated successful use of computer-based voice visualizations to facilitate speech production and vocalization in children with and without ASD/SPD. While prior work has focused on increasing frequency of speech-like vocalizations or accuracy of speech sound production, we believe that there is a potential new direction of research; exploration of real-time visualizations to shape multisyllabic speech. We developed VocSyl, a real-time voice visualization system with extensible architecture. Rather than building visualizations based on what "engineers" may think is needed, we designed VocSyl using the Task Centered User Interface Design (TCUID) methodology from the beginning to the end of the design process. Children with ASD and SPD, targeted users of the software, were directly involved in the development process, thus allowing us to focus on what these children demonstrate they require. This chapter presents the VocSyl system and architecture, the results of our TCUID design cycle, as well as design guidelines for future work with children with ASD and SPD.

9.1 Introduction

The process of developing language is "a unique characteristic of human behavior…[that] contributes in a major way to human thought and reasoning" (Lovaas, 1977). Moreover, the ability to combine syllables is a critical milestone in speech development (Highman et al., 2008; Tager-Flusberg et al., 2009), and is key to optimizing language growth. However, some children with **A**utistic **S**pectrum **D**isorders (**ASD**)[1] and **Sp**eech **D**elays (**SPD**) do not develop language on their own from the input typically presented to them (i.e., via the auditory modality (MacWhinney, 1998)). Hence, the natural

[1]The prevalence of ASD is estimated by the Center of Disease Control and Prevention (CDC) to be 1 in 150 children (Center for Disease Control and Prevention, CDC, 2007)

development of social behavior, such as spoken communication, is significantly disrupted. This results in detrimental effects on many aspects of their lives (Highman et al., 2008; Tager-Flusberg et al., 2009). We therefore sought alternative forms of presentation and feedback to facilitate their learning to communicate. Specifically, we aim to develop solutions to help teach multisyllabic word production; either as word combinations (e.g., "more juice") or as individual multisyllabic words (e.g., "banana").

Visual information has emerged as a critical form of support for children with ASD due, in part, to documented strengths of this form of processing information (Minshew et al., 1997; Peterson et al., 1995). Further, the use of technology has the potential to alleviate some apprehension experienced by many ASD and SPD children when interacting with people (Baskett, 1996). Given the success of HCI researchers in using visualization to encourage vocalization (Hailpern et al., 2009b), we strongly believe that voice visualizations on the computer have the potential to provide teachers and therapists with new techniques to complement and/or supplement existing approaches.

To develop these software tools we formed a team from Computer Science, Speech and Hearing Science and Special Education, and employed a **T**ask **C**entered **U**ser **I**nterface **D**esign (**TCUID**) methodology (Lewis and Rieman, 1993), in which subjects drawn from the target populations were involved with software development throughout the design phase. In effect, children with ASD and SPD became part of the development team. Because children with ASD and SPD have difficulties communicating, we examined their preferences as well as their interactions with the researcher and technology. To further inform our design, we included children without speech-language delays to provide explicit verbal feedback on our system. Children without delays can provide more explicit verbal feedback about the software, that can further guide the software development.

We begin with a review of the literature at the intersection of HCI, speech therapy, and ASD/SPD. Using prior work to guide our system, we next present the system and architectural features of VocSyl and describe how we ensured flexible and rich real-time visualizations. We then discuss the *TCUID Study*, including both the methodology used and the design changes made to our software based on the interactions in the design process. Our findings are summarized in a set of *Design Guidelines* for future research targeting multisyllabic speech production. Finally, we conclude with a discussion of an ongoing intervention study to examine the impact of our software.

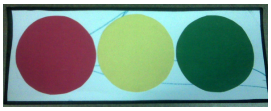


Figure 9.1: *Pacing Board for 3 Syllable Word*
Paper, with three construction paper circles glued down.

9.2 Background & Literature Review

For all children, multisyllabic speech represents a critical milestone – aiding the communication of more sophisticated concepts through phrase development and phonologically complex words (Velleman, 2002). Challenges with syllable combinations (within or across words) can impact a variety of children, including late-talking toddlers and children with apraxia of speech (Highman et al., 2008; Tager-Flusberg et al., 2009). In addition, there is indication that one out of every three children with ASD do not develop functional speech at all[2] (Bryson, 1996). Difficulties in multisyllabic productions can persist and result in other phonological difficulties later in life (Gallagher et al., 1996; Preston and Edwards, 2007). Given the importance of multisyllabic productions, it is striking how scant the literature remains in regard to related interventions and their effectiveness. The remainder of this section will focus explicitly on speech interventions, with a focus on children with ASD and more specific speech impairments. We focus on these two populations due to evidence of their difficulties in developing intelligible spoken language and evidence of the effectiveness of visual aids.

9.2.1 Therapy for Children with Autism & Speech Delays

Many of the characteristic difficulties experienced by children with ASD revolve around communication, empathy, social functioning, and expression. Since the 1960s, Ivar Lovaas' pioneering and successful approach of "applied behavior analysis," or ABA, has been one of the main methods[3] to teach communication and social skills (Lovaas, 1977). However, frequent therapy sessions that require sustained attention and intense human-to-human contact can produce anxiety, which can cause difficulty for both practitioners and children with ASD (Kanner, 1943). This motivated us to explore technology based solutions (which have the potential to minimize anxiety (Baskett, 1996)) that improve language skill, which may increase the willingness of children to participate in the therapy.

[2] Although speech is not the only means to successful communication, its development is consistent with other forms of communication, such as signs or picture symbols, and usually remains a relevant and meaningful goal (Millar et al., 2006)

[3] Other forms of communication treatment (Greenspan and Wieder, 1997; Koegel et al., 1999; National Research Council, 2001; Woods and Wetherby, 2003) have been used to help develop social and communicative skills in children with ASD.

In addition, providing information through the visual modality has repeatedly been documented as a useful treatment approach for children with ASD (Minshew et al., 1997; Peterson et al., 1995). For example, tools like the Picture Exchange System (Bondy and Frost, 2001) and visual schedules (Dettmer et al., 2000) have been widely adopted to support interaction and day-to-day functioning. However, the use of visual feedback to provide real-time acoustic information about vocal productions for children with ASD is rare[4]. In addition, few if any interventions for children with ASD have focused explicitly on multisyllabic productions. In contrast, a few treatment studies targeting multisyllabic productions have been published in relation to children with specific speech-language delays (e.g., (Strand et al., 2006; Yoder et al., 2005)). The most relevant approach is the use of a static visual display known as a pacing board (Velleman, 2002). The pacing board, represented in Figure 9.1, provides a graphic representation of individual syllables via circles. As a word is practiced, the clinician and child touch each circle as its corresponding syllable is produced. However, the pacing board is neither dynamic, nor does it provide acoustic information about the child's or clinician's vocal productions.

9.2.2 Technology Research on Autism and Speech Delays

Computers have been shown to be an important platform for interventions for children with ASD. Researchers have examined the role of technology in early diagnosis (Kientz et al., 2007), teaching human-to-human interaction to high-functioning children with ASD (Kerr et al., 2002; Tartaro and Cassell, 2008) and reducing the apprehension caused by human-to-human interaction in play (Michaud and Theberge-Turmel, 2002). Morris et al., have examined the customizability of computer systems to provide a more inviting learning environment for children with ASD (Morris et al., 2010). However, the majority of this research does not focus on speech acquisition and speech skills.

HCI research (Bergstrom and Karahalios, 2007c; Fell et al., 2006; IBM, 1997) and technology-based behavioral research outside the ASD community (Barry, 1989; Shuster and Ruscello, 1992) have illustrated the use of computer solutions in the context of speech therapy and communication. Within the ASD community, Hoque has explored sentence practice with higher functioning individuals with ASD through the use of games (Hoque et al., 2009).

Nonetheless, no prior work could be found that examined the role of technology in teaching multisyllabic speech. One potential reason is that many researchers using technology focus on lower and higher skill levels: sentences for high-functioning children (Hoque et al., 2009) or vocalizations for non verbal functioning children (Hailpern et al., 2009b). Because of limitations in speech recognition software

[4]SpeechViewer(IBM, 1997) and VisiPitch(KayPentax, 2008) are noteworthy exceptions.

(Nakagawa, 2002; Strik and Cucchiarini, 1999), forms of speech detection are limited, especially for individuals with poor diction. This poses a unique challenge. As a result, little work has been done to aid speech skills for individuals with ASD and SPD with severe speech deficiencies.

9.3 Scope & Motivation

Our research explores the process of building computer systems to aid in shaping multisyllabic speech in children with ASD and SPD, specifically by including these children, who have limited speech-language skills, in the design process. Because this design space is so large, and teaching multi-syllabic speech to children with ASD/SPD using technology is a novel interaction, launching straight into a validation study of our software would be premature (see Future Work for ongoing studies). Further, prior work has repeatedly shown that lessons learned from TCUID, loosely structured studies and case studies[5] are an important contribution (Edwards et al., 2003; Consolvo et al., 2006). They improve development time, usability, and help designers better understand context and users (Vredenburg et al., 2002). In this spirit, we wish to share our findings to help others working in a similar domain. Given the novelty of this design space, other designers can greatly benefit from the numerous technical and user-centered lessons learned, which we were only able to uncover through a TCUID investigation.

Our main contributions are the documentation of the TCUID process, the resulting design and architecture of VocSyl, and the articulation of design guidelines for future software development that shapes complex speech. Our research bridges the gap between Hailpern's work (Hailpern et al., 2009b) on encouraging vocalization in children and Hoque's work (Hoque et al., 2009) on providing a game to practice full sentence production. Note that teaching perfect speech or precise phonemes is outside the scope of this research: our software focuses on facilitating prosody and syllables. We employed well accepted **D**igital **S**ignal **P**rocessing **(DSP)** algorithms. While pitch detection and syllable detection[6] were not always 100% accurate(Nakagawa, 2002; Strik and Cucchiarini, 1999), our implementations were consistent across users. As this work focuses on user interaction, designing new or improving existing DSP algorithms (pitch, volume, or syllable detection) is outside of the scope of this work.

[5] Even without explicit validation studies.

[6] The algorithm used for syllable detection worked well for words with hard consonant breaks between syllables, but had increasing difficulty during words with soft syllable breaks. It addition ambient room noise would occasionally cause an extra syllable to appear.

9.4 VocSyl Software Architecture and System

VocSyl[7] is a Java based, real-time audio visualization system for use within a clinical speech-therapy setting. It utilizes the Processing API (Fry and Reas, 2007) to render graphics. To teach multisyllabic speech production, we visualize changes in syllables, timing (speed), tone (pitch) and emphasis (volume). Figure 9.2 illustrates four of the many permutations of VocSyl visualizations designed during this research.

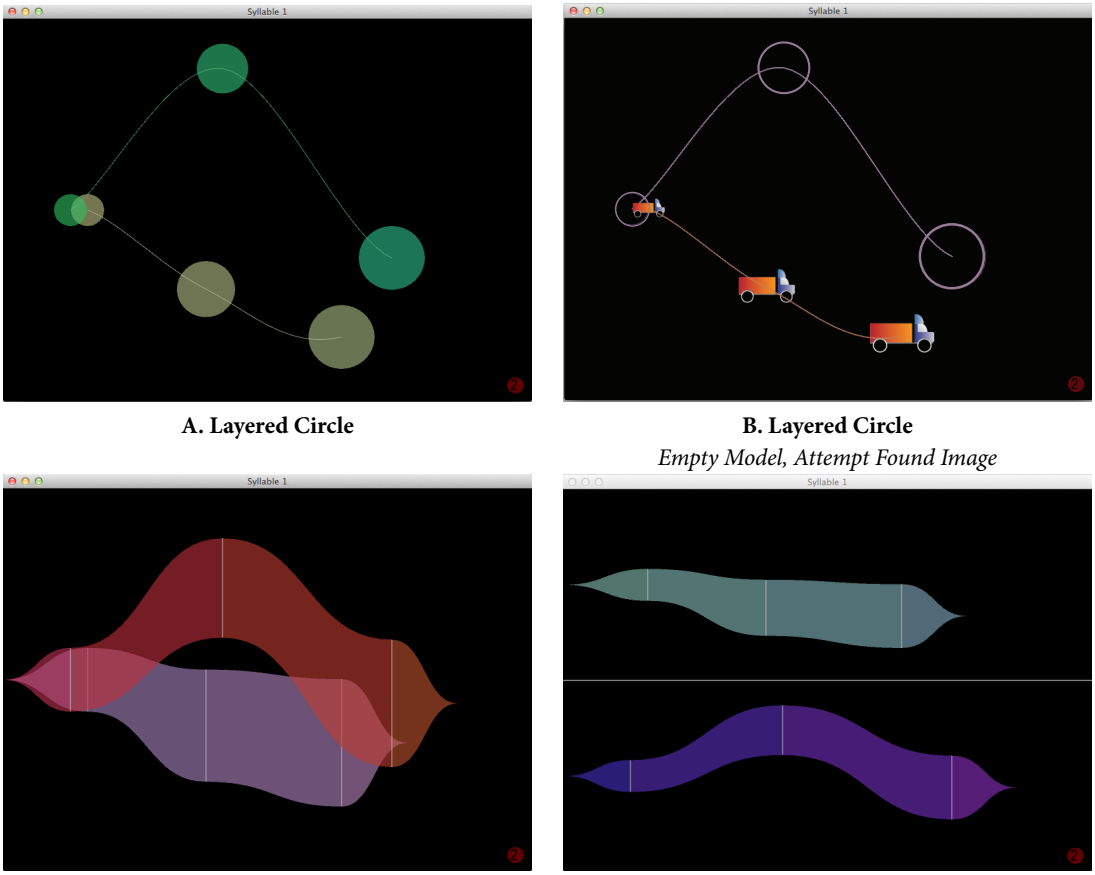
Research software is often designed and built with little consideration for the software's use beyond the conclusion of the experiment. This approach is effective for studying how humans interact with software, but it often does not yield software that is usable outside of a research setting. In contrast, we built VocSyl (and its visualizations) to be an extensible software architecture that could be used on a regular basis in a clinical setting.

We now highlight the target use of VocSyl and its visualizations. Because the VocSyl system itself is a major contribution of this work, we also highlight three key technological features: facilitating multiple complex real-time audio analyses, the audio processing architecture, and how VocSyl's design explicitly handles third-party development.

9.4.1 Use Case

VocSyl is based on a common form of behavioral training called **A**pplied **B**ehavior **A**nalysis, or **ABA** (Baer et al., 1968). The clinician first says a word. This prompt is called a **model**. The clinician then waits for a response from the subject. We term this response an **attempt**. If the attempt approximates the model, the subject is given a reward (e.g., food, verbal praise or a toy). If the attempt does not match the model, the trial is repeated. Within this communication treatment style, the prompt may be to say a specific work (e.g., "say cookie") or may be asking a question about an object (e.g., "what is this?"). When the clinician says a target word or phrase, that auditory information is fleeting and does not persist. Therefore, if a child says a word incorrectly, there is no lasting product presented to the child for comparison (aside from playing back a recording of their voice). Even the use of a mirror (a traditional speech therapy visual aid) (Coletto, 2009) is by nature temporal, and there is no representation of the target word after it is said.

[7] The name VocSyl is an amalgam of the words *Voc*alization and *Syll*able, acknowledging its target use case. VocSyl is pronounced as Voxel (/ *vaksl* /) as a nod to computer graphics and the visual aspects of the interface.



A. Layered Circle

B. Layered Circle
Empty Model, Attempt Found Image

C. Layered Envelopes

D. Split Screen Envelopes

Figure 9.2: *Four Examples of VocSyl Visualizations*
These examples are of two utterances of the word "basketball," the first where each syllable is said in descending pitch, and the second where the second syllable is said at a higher pitch than the first and third syllables.

VocSyl aims to present a visual representation (see Figure 9.2) of the vocal features of a word attempt. When the clinician prompts the child with a word, a visual representation of the uttered word appears in near real time on a screen. When it is the child's turn to attempt the word, their utterance is "drawn" in real time next to, or on top of, the visualization of the model. As a result of VocSyl's persistent visualization of the auditory features, the child and clinician can visually compare the differences between the model and the attempt.

9.4.2 Real-Time Multisyllabic Speech Visualization

This visualization has two basic **styles** loosely based on the pacing board commonly used in speech therapy (Figure 9.1). We term the style *circle* to refer to the shapes in Figure 9.2 A and B (where each syllable is an individual circle). We term the style *envelope* to refer to the shape in Figure 9.2 C and D (where each syllable is a different segment separated by a vertical white line). The circle visualization can be modified. The thin line smoothly connecting the circles together (Figure 9.2 A and B) can also be turned on and off. In addition, the visualization of the model has the option to be an empty circle (simulating a target to hit and match as in Figure 9.2 B).

This VocSyl visualization dynamically captures and shows four key elements of each utterance produced. *Syllables* are represented by discrete elements on the screen (envelope or circle). *Pitch or tonal changes* (relative to each user's starting pitch) are illustrated on the y-axis[8]. *Emphasis* or stress is represented by envelope or circle size (thickness or diameter respectively). *Pacing/timing* is represented on the x-axis.

9.4.3 Real-Time Event Based Audio Analysis

Systems which rely upon real-time visualizations, like VocSyl, require both visual rendering and audio processing. Because synchronicity is central to real-time visualizations, if one element of this system (audio analysis or visualization rendering) dominates the applications' execution, the whole system suffers. Thus, issues such as deadlocking due to rendering, processing, or analysis can prevent such systems from functioning in real-time (a critical component of VocSyl). Therefore we wanted to ensure that VocSyl would be highly scalable to multiple concurrent visual and auditory components. Further, the connections between audio analysis and visualization should be loosely coupled so that the system does not deadlock or lag.

[8] We used relative change because the voice of an adult is lower than that of a child. Relative change in cents is a linear delta between starting pitch and current pitch (unlike Hz which is logarithmic).

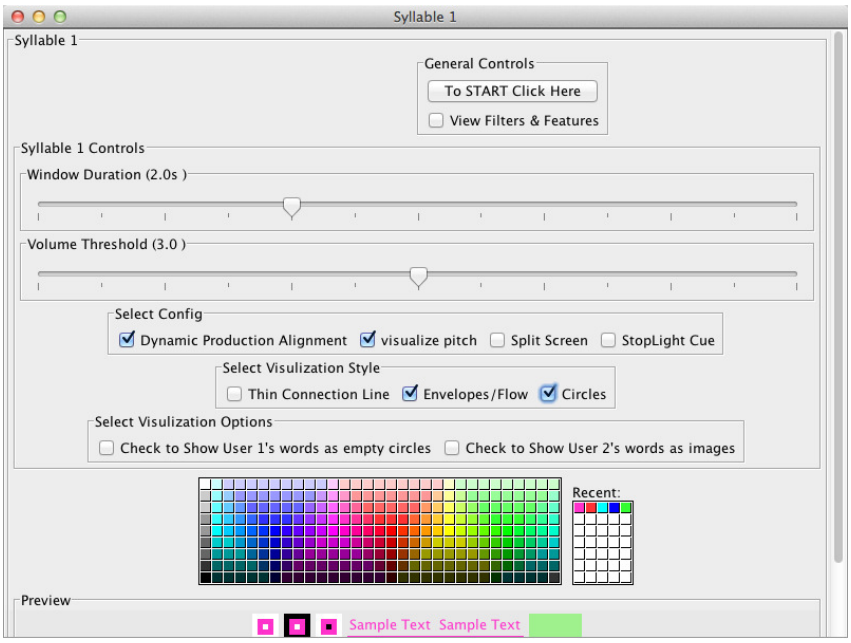


Figure 9.3: *VocSyl Visualization Configuration Window*
Properties of the visualization can be changed in real time (e.g., color, duration of the visualization window)

Thus, VocSyl follows the *Publish/Subscribe* design pattern (Gamma, 1995). Rather than a linear series of functions to execute, each **component** (visualization, microphones, audio feature extraction, etc.) runs autonomously on separate threads. Each component communicates with the others via ad hoc data passing. Think of the components of VocSyl as a series of radio stations and people listening to the radio. A "station" (e.g., a microphone) can broadcast (e.g., an audio sample) to anyone (e.g., volume extractor or pitch extractor) who has her radio "tuned" to that station. In addition, each listener has his own broadcasting station. Therefore, each radio listener can also broadcast her own radio signal (e.g., the volume and/or pitch of the sound sample) to anyone listening.

Similar to the radio station/listener analogy, the components of VocSyl are loosely coupled. No component needs know what the network topology is, or who is listening to it. No one component's processing or analysis of information will slow or delay the system. To counteract race conditions and asynchrony, all messages passed in VocSyl have timestamps. Another added benefit of this design is its scalability. For relatively small systems, Publish/Subscribe architectures scale better (as more components are added) compared to client-server based designs (Gamma, 1995). As a result of leveraging this design architecture, VocSyl and its components can be complex while maintaining smooth real-time visualizations.

Type	Name	Description
Mic	Microphone	Pulls Audio from internal or USB Microphone
Mic	WAV Audio	Pulls Audio from WAV file for debugging
Filters	Soft BPF	Configurable Band Pass Filter to gradually remove signal above/below a frequency (HZ) threshold
Filters	Hard BPF	Configurable Band Pass Filter to cut off signal above/below a frequency (HZ) threshold
Filters	HPF	Configurable High Pass Filter to remove signal below a frequency (HZ) threshold
Feature	Volume	Volume of current audio sample based on amplitude
Feature	dB Volume	Volume of an audio sample (Decibels)
Feature	Pitch	Pitch of audio, using a sliding window, calculated using a Cepstrum Plot (Roads et al., 1996)
Feature	Syllable	Detects when a Syllable occurs; alerts listeners using Mermelstein's algorithm(Mermelstein, 1975)

Table 9.1: *Mics, Filters and Features built for VocSyl*

9.4.4 Audio Processing Components in VocSyl

All components can potentially publish (broadcast) and subscribe (receive). However, the content of those broadcasts can vary. We categorize components as being one of three types:

- **Mic:** broadcasting an audio signal *e.g., audio from a physical microphone or audio from a WAV file*
- **Filters:** receiving an audio signal and broadcasting an audio signal *e.g., band-pass filter or low-pass filter*
- **Features:** receiving an audio signal and broadcasting a set of extracted value(s) from that signal *e.g., pitch or volume*

Table 9.1 lists all of the components (and their algorithms) built for VocSyl and describes their implementation. Based on the VocSyl architecture, many of the *Filters* and *Features* use each other to analyze and process audio. Due to the limitations in speech recognition software, especially for individuals with poor diction (Nakagawa, 2002; Strik and Cucchiarini, 1999), we did not build any *Features* into VocSyl that extract higher level audio properties (e.g. phonemes). Rather, the main

Filters and *Features* of VocSyl revolve around aspects of prosody: syllables (building blocks of words), pitch (intonation), time (tempo) and volume (emphasis). However, adding new components is a simple process, and is covered in the next subsection.

9.4.5 Plugin Architecture for Expandability

VocSyl's modular architecture allows the system to be highly extensible. Developers can easily add new *Mics* , *Filters* , *Features* or visualizations by building on top of the existing components built for this research. For example, a new visualization can be created in roughly 100 lines of code that makes use of the pre-built components that detect volume, pitch and syllables. While the specific visualizations used in this research are more complex (and required more code) than this simple 100 line example, our framework allows future visualization developers to focus on *how* to represent the vocal properties, and not the complexity of the larger VocSyl system. For visualizations, *Mics* , *Filters* or *Features* , developers need only follow the specified Abstract Class/Interface specified, and they will find that all the integration, broadcasting, and receiving has already been built in.

VocSyl also takes full advantage of the Factory Design Pattern (Gamma, 1995). As a result, if a user wanted to build a new visualization or component, VocSyl automatically detects it, registers it, and it becomes part of the VocSyl system. For visualizations and *Mics* , VocSyl automatically adds them to the VocSyl menu interface as though they "were always there." Likewise, if a user designs a new pitch algorithm or adds phoneme detection, they only need to write the one file and any component/visualization can now use it.

A further benefit that leverages the dynamic component detection is the ability to auto-generate configuration screens (Figure 9.3). Each component and visualization can specify properties that can be changed by a user. These properties are also auto-detected, and appear at the appropriate location in VocSyl. All the options displayed in Figure 9.3 are auto-generated, allowing properties to be added and hidden from view depending on what the user clicks and configures. This is another way in which developers can focus on their components without worrying about end user configuration.

Name	Gender	Age	Diagnosis
Sean	M	4	SPD
Tina	F	8	SPD *(SmithÐMagenis)*
Zev	M	10	ASD
Frank	M	14	ASD + Downs
Cara	F	5	Typical
Heather	F	6	Typical
Emma	F	5	Typical
Amy	F	6	Typical

Table 9.2: *Demographics*
Names were changed to protect identity, but gender was maintained.

9.5 Task Centered User Interface Design

Once the initial VocSyl software was completed, we began TCUID. Over a 6-month period, we examined how the children interacted with VocSyl, when they demonstrated preferences, and most importantly, the challenges they faced when interacting with voice visualizations that targeted multisyllabic speech. **It is important to note that our goal was not to design one "perfect" visualization, but rather to design a robust suite of visualizations, as the challenges, needs and preferences of each child will vary.**

9.5.1 Participants

Three groups of children were recruited for the TCUID. Two children with ASD were recruited from a school for children with disabilities. Two children with SPD were recruited from the local community. Four neurologically typical children [9] were recruited from a local preschool and child care program within a major university. Children were recommended by teachers and therapists and given an initial screening involving Part 1A of the *Communicative Development Inventory (CDI)* (Fenson et al., 2002) and the Words Sequence section of the *Verbal Motor Production Assessment for Children* (VMPAC) (Hayden, 1999). Descriptive information about individual children across all groups is provided in Table 9.2. All children were past the age of 18-24 months during which multi-word productions typically emerge. The typically-developing children produced and understood all 396 words on the CDI, whereas the two groups with speech-language impairments ranged from 23-308 words. Similarly, all children in the typical group correctly produced all 13 syllables from the Words Sequences on the VMPAC, whereas the other two groups ranged from 0-7 in terms of correct syllables.

[9] Neurologically typical children were recruited because they can verbally articulate their preferences and discuss their interests and confusion with the software.

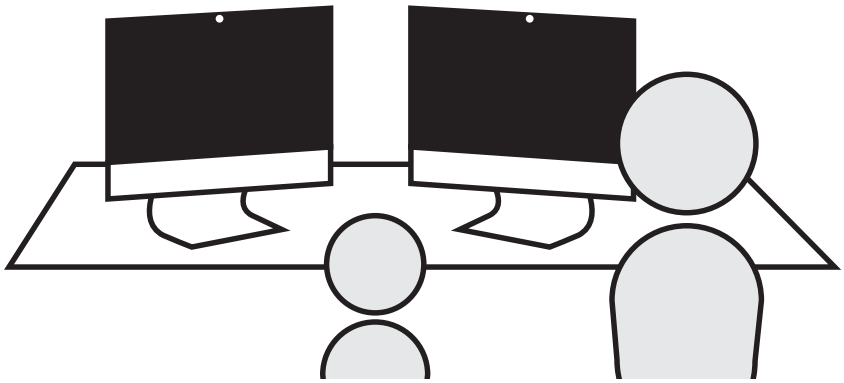


Figure 9.4: *VocSyl TCUID Setup*
The child would sit between the two computers while the researcher sat on the right side of the child. The child's chair would be a non-swivel chair.

9.5.2 Methods and TCUID Protocol

Participation in our TCUID began with an initial speech-language screening performed by a faculty member or a masters degree student in Speech & Hearing Science. Once the child was deemed eligible according to the inclusion criteria, he/she participated in six study sessions. Study sessions were separated by approximately 7-14 days (based on the availability of the child and completion of VocSyl software modifications). Sessions lasted about 15 minutes. Sessions were conducted in an unused classroom at the child's school or in the Computer Science department at a major university.

9.5.2.1 Protocol

At the start of each session, the child would sit in front of two computers running VocSyl (Figure 9.4) and the researcher would explain the protocol in age appropriate vocabulary. It is important to note that nowhere in the explanation did the researcher say, *"our visualizations should be touching."* We omitted this information to determine if the participants would find the graphical alignment of the visualizations to be naturally appealing.

Following the explanation, the researcher would then conduct four exposures. An *exposure* refers to a unique pairing of two visualization permutations (one on the left computer and one on the right computer). During the exposure, the researcher would run through a simplified ABA (Baer et al., 1968) protocol. This process would begin with the researcher randomly choosing one computer and saying, *"Let's start over here."* The researcher would then say a target word (e.g., *"Monkey"*), while

VocSyl visualized in real time. Then, the researcher would encourage the child to try (e.g., *"Now you say monkey"*). The child would then try to repeat the word while VocSyl drew their utterance. Regardless of the outcome (accurate or not accurate), the researcher would provide verbal word encouragement (e.g., *"Good Try!"*). This process would be repeated using the second computerÕs visualization. When both computers had the visualizations rendered, the researcher would ask the child, *"Which one did you like better?"* The selected visualization would be repeated. For children without ASD or SPD, the researcher would also ask the children why they chose the particular visualization. During each exposure, one word was used. No words were repeated across exposures or sessions.

9.5.2.2 Word Selection

Target words were selected to challenge childrenÕs productive abilities and thus emulate the intended purpose of VocSyl. For the children in the ASD and SPD groups, target words were selected based on results from the CDI (Fenson et al., 2002). Specifically, the examiner chose multisyllabic words that the child reportedly 'understands but does not say.' The words selected were either two- or three-syllable words. Given that the children from the typical group could reportedly say all words listed on the CDI, as would be expected based on their age, three- or four-syllable nonsense (selected from (Dollaghan and Campbell, 1998)) and low-frequency vocabulary words (published in (Mahurin-Smith, 2010)) were used instead. Overall, if children *could* imitate the words (even matching syllables or intonation) it suggests the potential applications of VocSyl within a speech therapy context.

9.5.2.3 Technology

The TCUID experiment utilized two 23" or 21" iMacs. For children with ASD or SPD, we used a Phoenix Duet USB Microphone with each computer. For typical children, we used the internal iMac microphones. Sessions were videotaped using SilverBack software which records both the screen and live video (using the iMac iSight camera). All video was recorded with parental approval.

9.5.3 TCUID Iteration Results: Six Rounds

We will now present the observations from the initial speech-language assessment and all six rounds of our TCUID process. For each round we briefly describe what the round was "examining," followed by

observations[10] of the typical children, then the ASD/SPD children. Round one examined the impact of visualization style (e.g., circle vs. envelope or solid circles vs. empty-circle). Rounds 2-6 each examined specific features of our visualizations by comparing a visualization with a feature turned "on" to the same visualization with the same feature turned "off." Decisions on which features to bring forward to future rounds were made across all children for consistency. Recall that we refer to the clinician's prompt/example of the word as a **model** and the child's response as the **attempt.**

While we do include the raw preference counts in Table 9.3, due to the small sample size, conducing traditional statistical comparisons is not done due to limited statistical power. These values are included for illustrative purposes only.

9.5.3.1 Initial Speech Assessment Observations

While the assessment session was not technically a "data gathering session," we noted that children with ASD and SPD were extremely quiet. They were willing to participate, to varying degrees, but their speech volume was low, and we realized the internal microphones on the iMacs would not pick up their vocalizations. We therefore modified our protocol to include the use of USB microphones that could be moved closer, or given to these children to hold.

9.5.3.2 Round 1 - Style

During the first round we examined the style of visualizations. For each exposure, we contrasted one of visualization style with another (e.g. Figure 9.2A with Figure 9.2C.) This included the envelope and the permutations of the circle style.

Since Round 1 was the children's first exposure, many seemed surprised and excited by the computer's response to their voice. Many of the typical children commented on the shape or alignment of the visualizations (e.g., Emma said, *"Mine is like green toothpaste!"* and Amy said, *"They're matching up with mine!"*). This type of response is one we saw repeated throughout the study.

When interacting with VocSyl, Zev often would lean into the iMac and touch the visualizations as they appeared on the screen. He clapped when the visualization appeared and said, "I did it!" during the session. Zev, Tina and Sean appeared very happy to be using the software and began smiling once we began using the software and it reacted to their voice. Tina often exclaimed, "Wow!" when the

[10] The qualitative observations and inferences in TCUID with children that cannot express their opinions are inherently subjective.

Round	A vs. B			A Count	B Count
1 *	Empty Circle	*vs*	Solid Circle	1	4
1 *	Envelope	*vs*	Solid Circle	4	2
1 *	Envelope	*vs*	Empty Circle	2	4
1 *	Circle+Line	*vs*	Circle Only	4	3
2 †	Split Screen	*vs*	Layered	12	19
3 †	no Pitch	*vs*	with Pitch	15	16
4	no DPA	*vs*	with DPA	16	16
5	no Images	*vs*	with Images	6	26
6 §	no Stoplight	*vs*	with Stoplight	16	12

Table 9.3: *VocSyl Preferences*
These values are included for illustrative purposes only.
∗Sean would not give preferences this round — § Tina was unable to participate in this round due to medical issues. — † During one exposure, one child did not give a preference (thus total preferences was 31 not 32)

visualization appeared. In addition, Tina also held the microphone and put it up to her mouth *only* when it was her turn. She would continue this behavior throughout the entire six rounds.

Overall, this round produced no consistent preference across or within our participants. All the children actively participated, attempted all the words presented to them, and all appeared to enjoy themselves.

9.5.3.3 Round 2 - Split Screen vs. Layered Screen

During this round, we examined the physical relationship of the visualization of the clinician's model to the visualization of the child's attempt. One computer in each exposure would randomly have the model visualized on the top of the screen with the attempt visualized on the bottom (Figure 9.2 D), while the other (with an identical visualization) would have the two visualizations overlaid (Figure 9.2 A, B or C).

Similar to Round 1, the verbal reactions of the typical children continued to be positive and analytical. These children regularly noticed patterns and relationships between the visualizations. Amy commented *"Boy, those ones are touching very good!"* indicating the fact that both the visualizations of the model and her attempt touching was a positive thing (despite not being mentioned in the protocol). While the reactions of Zev, Tina and Sean remained the same, Frank began to smile and readily engage with the researcher and VocSyl. While he was phonetically quite inaccurate, he matched the syllables and inflection of the researcher on each attempt. After his attempts, he would often smile as he looked at the computer screen. We did not see this response during Round 1.

The second round produced no preferential results, similar to the first round, with a few noteworthy exceptions. When we tested envelope split screen (Figure 9.2 D) versus envelope layered (Figure 9.2 C), all eight children preferred the envelope layered. However, for the other three exposures (which utilized variations on the circles), the preferences were split 50/50 between split screen and layered. This perfect split was also seen within the typical and ASD/SPD groups. Going forward, we opted to leave all visualizations layered.

9.5.3.4 Round 3 - Pitch

It could be argued that with four features being displayed (volume, syllable, pitch and timing), this may be "too much" for the children to process. For Round 3, we reduced the complexity of the visualization by examining the role of pitch in the visualizations. Visualizations without pitch (no variation in the y-axis) would render the visualizations closer in presentation to the pacing board. One computer in each exposure would have pitch "on" (allowing the y-axis to be used), while the other (with an identical visualization) would have pitch turned "off" (each syllable's y-offset was 0).[11]

In Round 3, the children generally preferred having pitch on. When asked to explain their preference, the typical children most often cited the relative position or alignment of the model and the attempt as their justification. Amy justified her preference by saying, "Because I like how it goes down and the circles are touching." Emma gave a similar rationale stating she chose one over the other, "because we have the same number [of circles]." Zev's reaction to pitch was also positive: he smiled, leaned closer to the computer and clapped when the pitch-on visualization came up. Given the positive reaction, we decided to continue using pitch in the visualization for the remaining three rounds.

A secondary observation in Round 3 is that all the typical children (and Sean from the SPD group) began asking to choose the color of the visualization. Color is a variable controlled by researchers and was not initially intended to be an option for the children. However, seeing that color preference was important to some children, we obliged. From this round forward, children changed the visualization colors during each exposure (though the color was always the same on both computers within any given exposure).

[11] Pitch *was* turned on during Rounds 1 and 2.

9.5.3.5 Round 4 - Dynamic Production Alignment

Over the first 3 rounds, we noticed that all of the typical children would often justify their preference, by commenting on how closely aligned the model and the attempt were. At times, however, the childrens' vocalizations would not directly align because their first syllable may have taken longer to say or their more explicit intake of breath would start the software. We hypothesized that many of the preference decisions were a result of the childÕs desire for alignment and matching the model. This would, in theory, be a positive observation in behavior in that it is the main goal of VocSyl.

To further explore the role of alignment in VocSyl, we created a feature called **D**ynamic **P**roduction **A**lignment (**DPA**). When DPA is activated, once the child has made an attempt to say the word, their vocalization would slide to the left or to the right in one or two seconds such that their first syllable would be directly on top of the researcher's first syllable. In theory, this would align the two productions, providing a more direct comparison between the model and the attempt. During Round 4, one computer in each exposure had DPA turned on, while the other (with an identical visualization) would have DPA turned off.

Remarkably, seven of the eight children took note of the movement in the DPA (either by verbally commenting, or visibly reacting to the animation). Notably from the ASD/ SPD group, Frank exclaimed "I like!" when he first saw the DPA, and Sean turned to his mother twice, saying, "it goes ma ma," and "it's moving mommy." Given this positive response, we elected to enable DPA in Rounds 5 and 6.

It is also worth noting that in Round 4, Tina explicitly asked (though with very limited articulation), "Do more," after we finished with the required exposures. So as not to bias her to future VocSyl exposures, we turned on an oscilloscope visualization which we had built previously for debugging VocSyl. Tina subsequently played with the oscilloscope twice, each lasting for about a minute. Once demonstrated, Tina said, "I will try," and she proceeded to say words into the oscilloscope for about 1 minute, concluding by saying, "My voice is on the 'puter." After playing on the iPad she asked to play on the computer again, in response to which the researcher brought up the oscilloscope visualization again.

9.5.3.6 Round 5 - Found Images

Images related to the childÕs interest are likely to spur engagement (Hailpern et al., 2009b; Morris et al., 2010). During this round, we examined the impact of replacing the abstract circle with found images (Figure 9.2B replaces the attempt's circles with trucks). One computer would randomly replace the

circles of the child's attempt with a found image, while the other (with an identical visualization) would use the original circles. *As there are no objects that could naturally replace the envelope visualization, we did not use it as one of the four exposures.*

We asked each typical child for a cartoon they liked, and parents/caregivers of the ASD/SPD children for preferences or interests. Overwhelmingly, children's preference was to have the images in VocSyl over the abstract shapes. Preference from the typical children was based on the presence of images (e.g., "I want Cinderella again! .. my path is Cinderella!" - Emma, "because Elmo, I like Elmo" - Cara). Preferences for the ASD/SPD children were equally based on the images. Tina attempted to count and point at the found images, Frank said "ohhhhhhh" with a look of immense pleasure on his face, while Zev would say "truck."

We observed that when children saw researchers turning on and off the images, both typical and SPD children started asking or pointing to the screen to choose their own images. Similar to the color selection that occurred in Round 3, an option to change the image was not intended by researchers, but was identified by the children.

9.5.3.7 Round 6 - Stoplight Cue

We noted that children had difficulty waiting to start their attempt. A few times, the children were so excited to say their word and see the visualization, that as soon as the researcher finished saying the model, they did not wait for VocSyl to start rendering their production (a process that takes seconds). We designed a visual cue to show when it is the child's turn. Specifically, after the researcher finished the model, a stop light appears and changes from red to yellow to green (over three seconds). When the stoplight rests on green, VocSyl begins listening to the attempt by the child.

Surprisingly, we had a very strong negative reaction from all children. When asked for a justification, typically they told us that they disliked waiting (e.g., "because I donÕt like to wait" - Amy, "when I do it with the stoplight, I have to wait" - Emma). Frank, rather than waiting, contented to repeat his attempt until the computer executed his visualization. This explicit delayed gratification and delayed interaction seemed to be less desirable, even though the children now knew exactly when to begin to interact.

In addition, two of the four typical children as well as Sean and Zev asked (verbally or by pointing) to have the found images turned back on in this round and Frank would move his hand horizontally in time with DPA aligning his attempt with the model. In addition, Frank spent almost a minute exploring his voice with the oscilloscope after the conclusion of the study.

9.5.4 General Observations

We observed that all the typical children commented or narrated, the visualization. Comments centered around the shape of the visualizations, who had "more circles," and when or where the visualizations touched. Cara, in particular, personified the circles as family members. She would label the "mommy" circle, the "daddy" circle, and the "baby" circle (based on size). Children with ASD and SPD, however, were not generally explicit or commented on the visualizations. This, may be a direct result of their linguistic and/or developmental challenges. **Even though this was not a treatment study, it is worth noting that the children with SPD and ASD we actively attempting words that they do not say.** More impressively, both children with SPD were saying the words well and the children with ASD made good attempts, almost always matching the syllables and pitch (and Frank regularly matching phonemes). Occasionally, Frank and Sean would repeat the same phonetic pattern across sessions (regardless of the model). We believe that this may be tied to words they found challenging, and rather than failing, they wanted to see a visualization react to their voice.

9.5.4.1 Positive and Enriching Experience

Across the board, children appeared to enjoy engaging with the computer and wanted to continue their interactions. We explicitly asked the children from the typical group whether they had fun and if they wanted to come back. Tina and Sean both asked to use the computer again (though these requests were "translated" by their parents), and Frank and Zev often talked into the microphone after exposures finished, continuing to look at the screen for a visualization of their voice.

This engagement is a important element of a positive learning environment, in that engagement is tied to learning (Skinner et al., 1990). Over the six rounds, the children became increasingly relaxed and engaged with VocSyl. All children (including the children with ASD/SPD) came into sessions smiling, rarely needing a prompt to sit down at the computers. Further, all the children attempted the words they were prompted with. This is particularly meaningful considering that all the words tested with the four children with ASD or SPD were words that they were not reported to use prior to starting the study. Though the phonetic accuracy varied by each child, attempts were made by all, and syllables and intonation were almost always correct. Even more significant was the reaction of Tina and Sean. Both SPD children tried quite hard to match the models and often said words almost, if not completely, perfectly. Tina's mother particularly commented on her daughter's ability to say the study's words, remarking how Tina never says these words at home *and* she was not currently enrolled in any

speech therapy. Tina's mother also asked for DVDs of her daughter's session to re-watch her daughter saying all the new vocabulary.

9.5.4.2 Physical Interactions

At the conclusion of each session, we allowed the children to play with an iPad. This was not an interaction to be explicitly tested as part of the TCUID, but rather a reward for participating. During our sessions, we noticed that the children (especially the ASD and SPD children) readily enjoyed playing with the iPad and smiled when they got the applications to respond to their touch. Based on this, we re-watched the recordings of the six rounds of data collection. During this, we noticed that all typical children regularly touched the screen of the iMacs to describe the visualizations they saw and Sean, Frank and Zev both attempted to interact with the visualizations on the screens of the iMacs.

9.6 Design Guidelines

From our observations of the TCUID, discussions with typical children and the existing literature, we have produced a set of guidelines for designing software to facilitate multisyllabic speech production. While some may appear "obvious," these guidelines presented are grounded in experimental observation rather than anecdotal knowledge.

R1 **Minimize Delay to Interaction:** When designing software for children, ensure that when the child wants to engage, and the software is ready to respond and delays are minimized. Thus children are engaged and stay interested with the interaction, the software, and learning.
As we saw with the stoplight condition, and the children's desire to keep interacting, the computer visualizations are a highly motivating media for vocalization. However, the more delays we build into the system, the more frustrating and confusing these interactions appear to be.

R2 **Real-Time is Fun:** By showing visualizations change in real time, children's attention remains with the software. They then continue to perform the task/activity. By ensuring real-time visualizations, we hopefully can encourage learning (Skinner et al., 1990).
Given that our tasks were well-structured, we were unsure if they would still encourage interaction (as compared to the free-form task of (Hailpern et al., 2009b)). Our observations appear to suggest that children truly enjoyed playing with the computer, as their requests to keep going, verbal or otherwise, continued even after the official session had concluded. As visualizations change and animated in real-time, children both

discussed the changes and moved in concert with the animations. This avenue of research is promising, and supports further exploration of how complex interactions can get while still encouraging skill development and engagement

R3 **Child Customization:** Interaction designers should support simple, uncluttered option menus for children to make choices in the interface themselves – ideally allow children to point to and touch what they want. This ensures that as the child's own preferences change (within and between sessions), so can the software. By designing easy to use interfaces, children can feel empowered and engaged with their interactions. By employing touch, non-verbal children can customize the software themselves.

The degree of customization children requested with VocSyl was an unexpected finding. Both typical and ASD/SPD children requested color changes or picture changes, as soon as they observed that the researcher was changing a setting (or with color, that there was a grid of colors on the settings panel). This was a natural and unprompted gesture seen in nearly every instance that the children greatly enjoyed. While the work of Morris (Morris et al., 2010) suggests that customization can be facilitated in higher functioning children, it is noteworthy that children here actively sought out customization in the VocSyl interface.

R4 **Dynamic Computer Correction:** When the target child is learning a complex skill (e.g., speech), the software can "smooth" their interactions by auto-correcting or auto-adjusting minor mistakes without negatively impacting child's interaction. This can allow the clinician to focus on the skills being targeted rather than trying to explain minor errors by the child (or the software).

We explicitly manipulated word attempts by children using DPA. Given the positive reaction to DPA, we believe that systems, such as VocSyl, need not be completely literal in their visualization. Computers provide a unique ability over non-dynamic tools (Figure 9.1) to correct tiny mistakes made by a child or a researcher in order to focus on the goal therapy. It appears that if changes made are obvious to the child (e.g., sliding of the production rather than a sudden jump), they are accepted and understood.

R5 **Robust Microphone Setup:** Systems should provide a robust setup to accommodate multiple voice levels, as children have different comfort levels with the use of their voice. As children are less sure of their abilities, they may be more hesitant to loudly engage. As their comfort level rises, so will their voice level.

We found that neurologically typical children were more willing to speak loudly than children with ASD/SPD. Internal computer microphones did not suffice. Even an external USB microphone may need to be placed very close to the child. A USB microphone also provides a physical device that the child can hold onto and direct their engagement towards.

R6 **Competence of the Child:** Designers can "raise the bar" on targeted tasks and make them more challenging by allowing the clinicians to adjust the "picky-ness" of the software for assessing correctness. Further, tasks asked of children should likewise be able to become more complex.
While one half of the children in our study had speech delays, they were capable of successful interactions with VocSyl, regardless of complexity. We therefore encourage designers to be cautious with respect to oversimplification of software. Take on fairly complicated projects in this direction, with the option to dial back features. Further, given the interactions observed with Frank and Zev we strongly believe that all the children fundamentally understood that their voice created and manipulated the visualization.

R7 **Physical Interaction:** Children want to touch everything. Touch is an easy-to-understand interaction. Therefore, design systems that not only respond to touch, but provide meaningful feedback for those interactions.
We observed physical interaction with the computer on multiple occasions. We believe that large screens, animations, and vibrant shapes prove an engaging visual experience. All but two of the children attempted to touch the computer screen to interact with the shapes or to comment on them. We strongly believe that encouraging physical interaction can have an impact on the visualization and can provide therapists with new techniques to further teach and shape vocalization.

9.7 VocSyl Touch

Following our unexpected observations about physical interactions with the iMacs, we have ported the VocSyl system to the iPad and iPhone due to this iOS version has many added benefits over its desktop implementation; portability, size, battery life, and easy distribution to homes and clinicians. Most notable is the addition of touch interactions. While the VocSyl Touch implementation visually looks and reacts the same as the desktop version, when a clinician or child touches the syllables in the visualization, VocSyl slightly animates that touch and plays back that segment of the production from a live audio recording.

9.8 Conclusion

The primary aim of this research was to design and build a software tool called VocSyl for use in speech therapy to encourage multisyllabic speech production in children with ASD and SPD. Our goal was to include children with ASD and SPD *in* the design process through TCUID so as to emphasize

building not simply what engineers thought was needed, but what the intended users demonstrated they wanted. This user-centered approach allowed us to design a software system that is both highly configurable and provides a platform for a meaningful exploration of multisyllabic speech productions. In addition to developing our software system, we also generated design guidelines for future research and software working with this population to teach speech skills.

Given the overwhelmingly positive response from children and parents to VocSyl, we believe that our software and computer based interventions have the potential to improve multisyllabic speech production in children with ASD and SPD. While this was *not* a treatment study, the children involved were actively saying words using VocSyl that they were not saying at home. This adds further support to the potential impact of computer based visualizations in speech therapy.

Chapter 10

Summary of Findings

This part of the dissertation presents the motivation, background and theory related to the use of visually and auditory feedback to encourage vocalization in non-verbal children with ASD. Based on the existing literature, we developed SIP or the Spoken Impact Project. SIP consists of three major thrusts of research; 1) A framework to facilitate video annotation, and an instantiation of said system called VCode and VData; 2) A^3, a set of coding guidelines that allow researchers to assess non-verbal subject interaction with computer based feedback; and 3) A detailed quantitative analysis of SIPS through the use of VCode, VData and A^3. Results from this analysis suggest that computer based feedback systems hold the potential to greatly impact the vocalization of non-verbal subjects. Based on our follow up Wizard of Oz Study, we believe that systems like SIPS can be further developed to not only encourage vocalization, but also teach verbal communication skills.

The implications of this thesis part are broad and far reaching. Results form our research on video annotation provide a framework for software designers to better meet the needs of researchers, clinicians, and to improve video annotation accuracy and quality. In addition, two tools that embody the annotation framework have been rereleased, with impact reaching world wide to a plethora of disciplines from Neurosicence to Computer Science.

The A^3 guidelines demonstrate how research from multiple domains can be brought together to create a system for analyzing the behavior of non-verbal subjects. By leveraging the rich body of literature, we were able to build a robust framework that leverages the theory and years of research already conducted. Based on the findings for inter coder reliability, the A^3 guidelines provide a reliable system to quantitatively assess subject behavior. A^3 used in conjunction with VCode and VData create a system that is reliable, fast, and accurate for assessing non verbal subject behavior.

Given the results from the SIP study, we believe that Audio and/or Visual Feedback can be used to encourage spontaneous speech-like vocalizations in low-functioning children with ASD. More specifically, CFS appears to encourage the rate of vocalizations that can be considered Speech-Like

(sounds that can be phonetically described), while generally having little impact on those vocalizations that we call Non-speech-Like (e.g. grunt, screech, etc). In other words, CFS has the potential to differentially impact (favorably) speech-like versus non-speech-like vocalizations. SIP also suggests that low-functioning children with ASD may have distinct and varied preference for forms/styles of feedback. As a result, individual customization may be necessary in future situations. Though the range of variation necessary is unknown, the final solution might include a suite of feedback styles that may be selected by the parent, clinician, or child.

With the positive results of our data, the encouraging messages of parents, and the potential impact demonstrated in the Wizard-of-Oz study, we believe that SIP-styled therapy is an exciting and viable method for encouraging speech and vocalization in low-functioning children with ASD. This research presents the first steps towards uncovering the area of using audio and visual feedback to encourage speech in low functioning children with autism. In other words, SIP is a starting point for future research.

In order to explore one aspect of the SIP-styled therapy, we constructed VocSyl for use in speech therapy to encourage multisyllabic speech production in children with ASD and SPD. Our goal was to include children with ASD and SPD *in* the design process through TCUID so as to emphasize building not simply what engineers thought was needed, but what the intended users demonstrated they wanted. While not a clinical test of the VocSyl system itself, we were able to uncover 7 key design requirements for future software design. Further, our TCUID process also uncovered the overwhelming participation of subjects to interact and attempt word productions. This is noteworthy in that all the words attempted were all new vocabulary. While the observations of researchers may be bias, parent observations were equally supportive, with one mother even requesting DVD copies of her daughter's sessions.

10.1 Limitations

During the SIP experiment, children participating were diagnosed with autism and had significant intellectual disabilities. Their attention to tasks was limited. Sometimes the subjects would appear highly engaged with a form of feedback, while other forms proved completely un-engaging. This often resulted in trial sessions of extremely short duration, as subjects would get up and move away from the computer. Duration of our trials had high variance, and reduction in observation time may have reduced statistical power of this study and ability for statistical tests to reach significance. We may not

have fully appreciated the positive effects of SIPS in this small study. However, we were able to observe numerous forms of feedback that garnered significant changes in SSLV.

Due to the small scale and low statistical power of this study, conclusions drawn are at best preliminary and we cannot conclude that audio/visual feedback will increase SSLV for every child with ASD. The purpose of this research analysis are not to provide definitive conclusions on the impact of CFS, but rather to suggest its potential as a new direction of HCI research and to highlight factors that researchers should take into consideration. Because of multiple testing and our decision not to make a statistical adjustment, it is possible that some of our significant findings may have been spurious. While these limitations are important to note, they do not greatly hamper the position or potential of the findings to shape future investigations. Further, based on our 5 single-subject studies, we believe our results are promising.

Given the small scale of the TCUID study, we cannot conclude that VocSyl will explicitly improve speech therapy and multisyllabic speech production (though qualitative observations do appear promising). However, this was not the aim of the current study. Prior to running a fully powered study, we aimed to ensure that the VocSyl and its interactions were designed carefully, with consideration of the end-user. To further this investigation, we are currently conducting a mixed method intervention study of 18 children with speech-language impairments enrolled in one of three conditions: intervention with VocSyl, traditional therapy with a pacing board, and a playgroup aimed at social interaction. The goal is to examine the impact of all three conditions on children's multisyllabic productions.

We also wish to highlight that there is a leap between producing SSLV, multisyllabic speech and real-world communication. Our studies to date have focused on encouraging a specific behavior, each of which is one component of functional communication. This work, in conjunction with the findings from other research, lays the groundwork for future exploration of this area.

10.2 Future Work

Given our encouraging results, there are many exciting areas of future work. One of the most immediate directions is adaptive feedback selection. Previously, researchers had to qualitatively assess which visualizations and forms of audio feedback were engaging to subjects. Future work might examine if a system could adaptively change forms of feedback by the subject's response via machine learning. This would not only ease the job of clinicians and researchers, but as preferences change and subjects satiate, such a system would be able to adapt.

We see the potential to test our approach with other populations or other target behaviors. Clearly, VocSyl is a first step. However, one additional avenue of exploration with VocSyl is phoneme production. Some of the children in our study have speech sound (phoneme) impairments. We believe that the system could be expanded easily to teach and shape phoneme acquisition, and are actively expanding VocSyl to target this. Another opportunity would be to explore the delivery of a SIP or VocSyl appliance. The investigation of a toy-like device could provide therapeutic play at home, as well as the practitioner's office. The implementation of VocSyl Touch begins this process, though other form factors should be explored.

Part II

Aphasia, Linguistics, and Conversation Log Analysis

Chapter 11

Introduction to Aphasia, Linguistics and Conversation Log Analysis

Receiving empathy and understanding are a constant need for those living with aphasia (Liechty and Heinzekehr, 2007). Aphasia is an acquired language disorder that impairs expressive and receptive language (both spoken and written)(Sarno, 1998). Over one million Americans are affected by aphasia (National Institute on Deafness and Other Communication Disorders, 2010). There is considerable variability in the effects of aphasia, but all individuals with aphasia display difficulty producing words accurately and with ease. Specific disruptions in language include dropping key words, substituting incorrect words for the target word, mixing up sounds or letters within words, describing items that are difficult to name, having difficulty organizing coherent strings of words, and making multiple attempts to correct such production errors.

To an outsider it may appear that an aphasic individual has poor cognitive function. However, the problem resides in the individual's receptive and expressive language, not their ability to think. Aphasia most profoundly affects the ability to communicate with others, whose lack of understanding and empathy has the potential to "erode the social bonds that give life meaning" (Liechty and Heinzekehr, 2007).

Our goal is to promote empathy and understanding of aphasia in unimpaired individuals. We made it possible for those without aphasia to metaphorically "walk in another's shoes" by interacting with a system which emulates the effects of aphasia through distortion of written text. This system has the potential to help family and friends of individuals with aphasia, and to serve as a training tool for physicians, nurses, and speech-language pathologists.

We introduce a novel system and model, called Aphasia Characteristics Emulation Software (ACES), that enables users to experience the speech-distorting effects of aphasia. This model adheres to the wealth of literature in language distortions resulting from aphasia. It was informed though a series of interviews and demonstrations (using an early prototype of ACES) with professionals and students from the field of aphasia and other language disorders.

The ACES system was designed to distort a user's Instant Messages (IMs) from the original message to one that appears like a message spoken by an individual with aphasia. Thus, the conversation that develops between the user and their IM partner has similar difficulties and hurdles to those experienced by an individual with aphasia. By experiencing these challenges firsthand, we hypothesize that users will have increased empathy, knowledge, and understanding of aphasia. Results from an evaluation of 64 participants indicate that ACES provides a rich experience that strongly increases understanding and empathy for aphasia.

The foremost contribution of our system is a demonstration of how existing literature about language distortions/errors due to cognitive impairment, age, or other language-based challenges can be extended to an interactive system that increases knowledge, empathy, and awareness. Our goal is not to perfectly emulate aphasia.

With this in mind, we conducted a a Turing Test study in which participants must distinguish samples of distorted text generated by a human from samples of text distorted by a computer. This work clearly demonstrated that ACES generates realistic aphasic distortions, thus validating the applicability of ACES as aphasic emulation software and supporting the feasibility of other language emulation software research.

Finally, we explore the potential of ACES as a tool to understand language and conversation quality changes. We conducted a 96-participant study with two major purposes. The first is providing Psychology researchers with a large corpus of conversations (with quality of conversation from both perspectives logged in addition to meta-data on what was intended to be sent). Second, and perhaps more importantly, is extending the original ACES work by demonstrating that our system can be used to understand how distortions impact linguistics and conversation quality. By exploring (at a high level) the conversation logs themselves, uncovering patterns in communication and how they relate to the frustration, clarity, and other dimensions of conversation quality, we can illustrate multiple future applications of the ACES approach.

This work, together, highlights the great potential of language emulation software; systems that sit at the intersection of human-computer interaction, psychology, and speech and hearing science. At this intersection, our technological solutions can greatly impact quality of life, quality of care, and knowledge of langue disorders while simultaneously developing new computer technologies that advance our understanding of how people interact with technology (and how to design new solutions for interaction).

Chapter 12

Review of Aphasia Literature

We describe aphasia, and existing research in the HCI community that relates to aphasia. Additional literature analysis is presented at the beginning of Chapter 13: Building & Designing Aphasia Characteristic Emulation Software, and Chapter 16: Write-it Do-it, providing additional related work pertaining specifically to the topics covered in those chapters.

12.1 Aphasia

Aphasia is a term used to describe an acquired language disorder that is caused by damage to the left or dominant hemisphere of the brain and impairs an individual's ability to produce and understand language in both written and spoken forms (Benson, 1979). The severity and pattern of aphasic symptoms vary, depending in part on the specific locations of brain damage. Clinical researchers have developed classification systems that identify different patterns or sub-types of aphasia. For example, diagnostic batteries (Goodglass et al., 2001; Shewan and Kertesz, 1980) based on the Boston classification system are designed to categorize an individual's aphasia symptoms as either a type of non-fluent aphasia Ê(Broca's, Transcortical Motor, Global) or fluent aphasia (Wernicke, Transcortical Sensory, Conduction, Anomic). Of particular interest to the goals of the current study is that all individuals with aphasia will display at least some difficulty with writing, and although writing may be more or less impaired than spoken language, the linguistic deficits in writing will be generally consistent with those of the person's spoken language (Benson, 1994). Recent research focusing on issues of treatment and functional recovery in aphasia(Chapey, 2001) has drawn attention to the need for clinical interventions to attend not only to the areas of deficits in the patient with aphasia, but also to the person's communicative and social systems more broadly. The research in this dissertation focuses on increasing empathy for those interacting with aphasics.

Some research, content, and text from this Chapter is reproduced from (Hailpern et al., 2011a; Hailpern et al., 2011b)

12.2 Computer Science Research on Aphasia

Researchers in the field of Human Computer interaction have largely focused the communication challenges pertaining to individuals with aphasia, in particular the use of visual images to aid in communications. Allen's research in 2007 and 2008 (Allen et al., 2007; Allen et al., 2008) has examined the role of image capturing mobile phones to allow individuals with aphasia to communicate by leveraging photos they have taken. This draws a striking parallel to Damen's research on visual storytelling to aid in communication (Daemen et al., 2010a). In many respects, the lessons learned of Allen and Damen's research come together in Al Mahmud's 2010 paper on XTag, in which the individuals location and images taken, are used to tell stories and recount activity (Al Mahmud et al., 2010).

Similar to the use of images to recount events, many HCI researchers have also examined the role of images in day-to-day communication. Boyd-Graber (Boyd-Graber et al., 2006a) examined PDA devices for word finding through images. Similarly, Nikolova explored visual interfaces for word finding (Nikolova et al., 2010; Nikolova et al., 2009) by using the "familiar physical interface"(Chao, 2001) as seen in products like Microsoft Bob (Microsoft Corporation, 1995) and MagicCap (General Magic, 1994). Ma et. al. explored the trade offs between images and iconography as a means for supporting image-based communication for individuals with aphasia (Ma et al., 2009). They found that images are just as functional as using iconography, expanding the possible images available to software developers. While non-visual, AVN (Al Mahmud and Martens, 2011) provides a tree-based word selection tool, to aid individuals with aphasia compose emails.

While much research has focused on facilitating communication of the individual, another branch of HCI research has explored facilitating speech therapy. Piper's pen based system allows individuals with aphasia to have aural words associated with images throughout he use of a paper-digital interface (Piper et al., 2010). Lewis et. al. has explored a more mobile perspective for word practice and speech therapy, by developing Banga, a mobile phone software system (Benjamin et al., 2008; Chandler et al., 2009). Moving further away from the visual and non-visual support tools in speech therapy, researchers in the field of artificial intelligence have explored the application of AI algorithms to support speech therapy for individuals with aphasia (Masson and Quiniou, 1990)[1].

Beyond the role of software in facilitating communication, HCI researches have explored the use of visual schedules in aiding the day-to-day activities of individuals with aphasia (Moffatt et al., 2004a), as well as visual cook-books(Tee et al., 2005). By studying the challenges that individuals with

[1] A full copy of this text was unable to be obtained by the author. The summary is based upon the published abstract.

aphasia experience when using software (McGrenere et al., 2003; Davies et al., 2004; Brett, 2009) and conducting participatory design studies(Moffatt et al., 2004a), new software developers can better develop and create technology to meet the unique needs of individuals with aphasia.

While the breath of HCI and Computer Science research on aphasia is quite large, little work has directly explored the use of software solutions to emulate the language distortions of aphasia, nor build empathy.

12.3 Other Emulation

In 1967, Weizenbaum et. al. intro ducted Eliza (Weizenbaum, 1967), a computer AI that attempted to emulate a Rogerian psycho-therapist. Four years later, Colby et. al. published PARRY[2] (Colby et al., 1971), and interactive computer model of paranoia. Colby originally intended PARRY to be used as a training tool by students (Boden, 2006), rather than allowing students to interact with actual individuals with paranoia. Both of these works illustrate how computer systems can be built to model and emulate a language perspective. Eliza, a psychotherapist, and PARRY, an individual with paranoia. However, neither of theses softwares allows the user to *experience* the perspective or disorder first hand (to build empathy). Rather, they are tools to ease the strain on professionals (Eliza) or provide training (PARRY).

To emulate the accessibility/readability of webpage for blind users, Takagi's built an emulator called Accessibility Designer (Takagi et al., 2004) that allows a web developer to see (through obfuscation) the readability of web content to a blind screen-reader user. The system obscured web content with shades of grey such that less accessible content is darker grey.

In a less "computer" based emulation, Tholen et. al. (Tholen et al., 2011) explored the effects of Dyslexia through the use of a Taboo-like game where participants would communicate with a list of words that they cannot use. While this study was not a computer emulation, it did explore the use of non-impaired individuals to better understand a disorder.

[2] PARRY passed a turing test experiment in 1972 (Colby et al., 1972).

Chapter 13

Building & Designing Aphasia Characteristic Emulation Software

Individuals with aphasia, an acquired communication disorder, constantly struggle against a world that does not understand them. This lack of empathy and understanding negatively impacts their quality of life. While aphasic individuals may appear to have lost cognitive functioning, their impairment relates to receptive and expressive language, not to thinking processes. We introduce a novel system and model, **A**phasia **C**haracteristics **E**mulation **S**oftware (**ACES**), enabling users (e.g., caregivers, speech therapists and family) to experience, firsthand, the communication-distorting effects of aphasia. By allowing neurologically typical individuals to "walk in another's shoes," we aim to increase patience, awareness and understanding. ACES was grounded in the communication science and psychological literature, and informed by an initial pilot study.

13.1 Related Empathy Work

We describe aphasia, empathy and how our work builds upon, and extends, the literature related to empathy and aphasia.

13.1.1 Empathy and Aphasia

Empathy is one of the fundamental underpinnings of interpersonal communication. It is an emotional response to the experiences of others, through which an empathetic person can understand and predict the feelings, emotions, and thoughts of others (Dymond, 1949; Stotland, 1969).

Non-technical approaches have been used to increase awareness and empathy in other situations by placing students in the "role" of an individual with an impairment. For example, students were tasked

Some research, content, and text from this Chapter is reproduced from (Hailpern et al., 2011a)

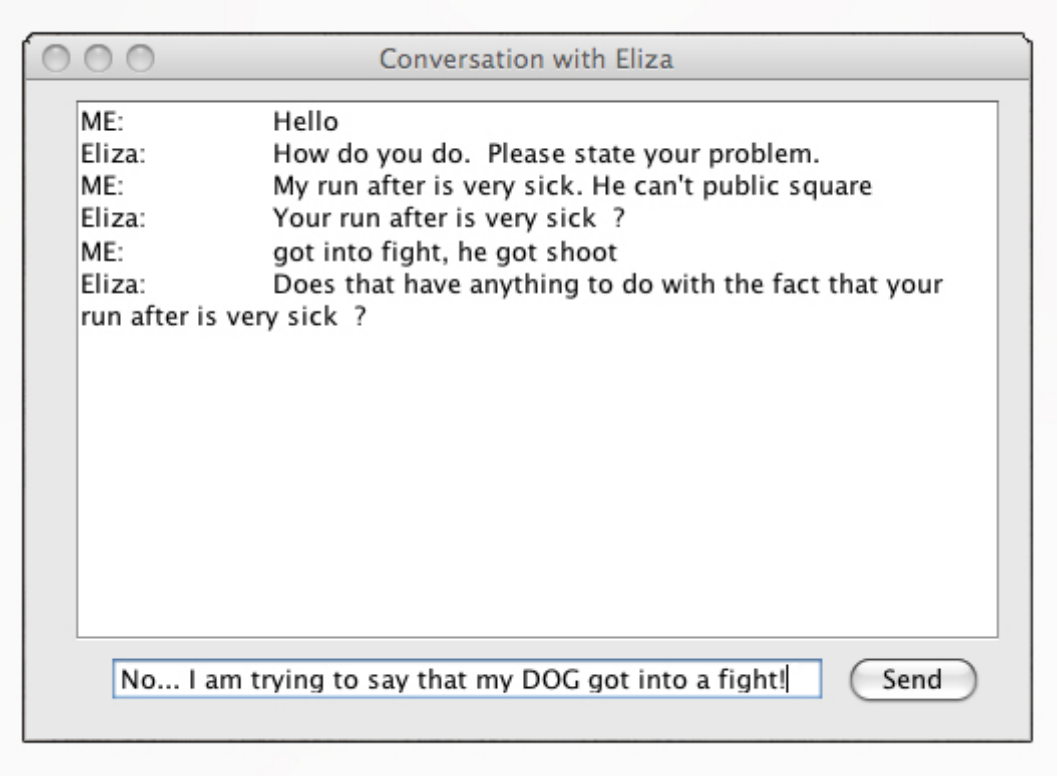


Figure 13.1: *Prototype Chat Window*
Notice how the participant's text is hard to decipher after it has been distorted by the Aphasia Model.

to spend a day in a wheelchair to develop an awareness of the challenges confronting a paraplegic (Doerksen, 2009). In another course, students were given Augmentative and Alternative Communication (AAC) devices (such as text-to-speech systems which assist those with impairments or restrictions on the production or comprehension of spoken or written language) to emulate "non-vocal" communication while performing tasks such as going to a coffee shop (Hengst, 2005).

If individuals relating to those with aphasia lack empathy and understanding, it can greatly reduce quality of life for aphasic individuals (Liechty and Heinzekehr, 2007). Often, family members can deny or underestimate the severity and presence of aphasic errors (Czvik, 1977). Further, in speech therapy, empathy is necessary to motivate the aphasic client, with motivation being one of the three key aspects of effective treatment (Shill, 1979). To date, research has shown that family member's ability to relate and empathize is based on how well they understand the distortions that their partners make (Furbacher and Wertz, 1983).

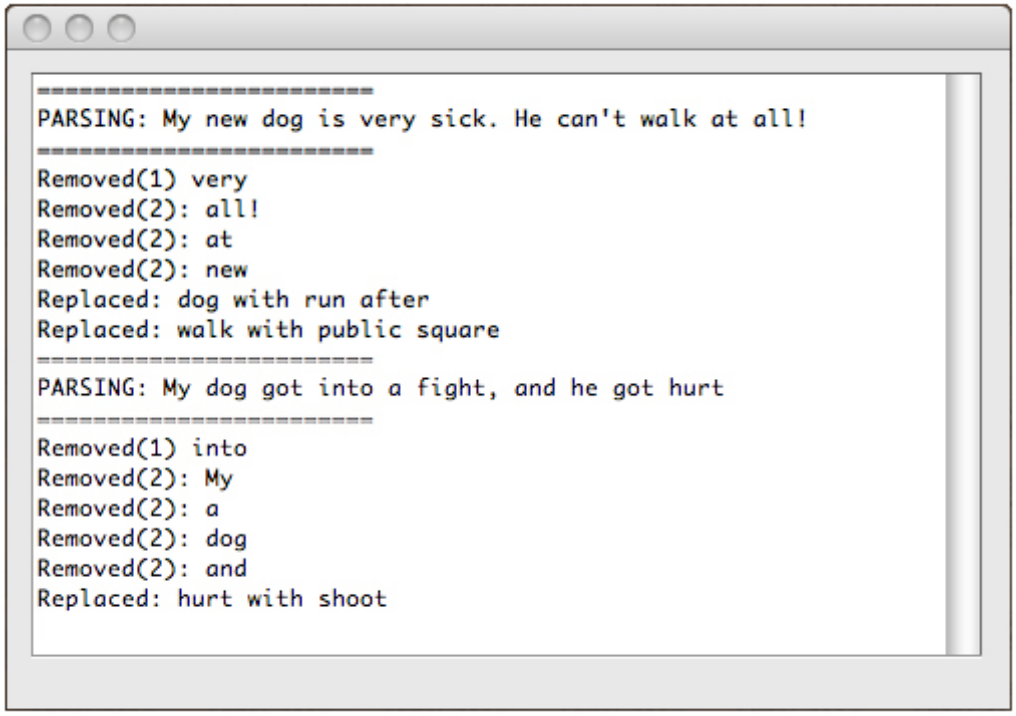


Figure 13.2: *Prototype Distortion Window*
Removed (1) is a dropped Function word, Removed (2) is a random word drop and Replaced is when a word is substituted.

13.1.2 Technology, Empathy and Aphasia

Technology has been used as a tool to enhance the functioning and communication of individuals with aphasia (Daemen et al., 2010b; McCall et al., 2009; van de Sandt-Koenderman et al., 2005) and other speech and language disorders (Kitzing et al., 2009; Mullennix and Stern, 2010). Specifically, work in the Human Computer Interaction (HCI) community has examined technological solutions to help with day planning (Moffatt et al., 2004b), communication (Boyd-Graber et al., 2006b), and increased access to web-content (Devlin and Unthank, 2006) for aphasic individuals. Most of this HCI research has focused on providing support for the individual, their communication, and providing awareness of their activities to those around them. Unlike work that focuses on developing empathic agents (McQuiggan and Lester, 2006), this work aims to increase empathy, knowledge, and understanding of aphasia for those who support aphasic individuals.

13.2 Early Prototype and Initial Investigation

Although there is a wealth of technology that aids individuals with aphasia, to our knowledge there is none that aims to increase empathy among those associated with aphasic individuals. Following the existing work in other domains (Doerksen, 2009), we hypothesized that if individuals were to experience the effects of aphasia first hand, they might better understand the disorder and become more empathetic to its challenges.

To this end, we built a prototype chat client that randomly distorted messages sent by its user using an aphasia language model. We chose IM, because it eliminates any bias from voice inflection and focuses on communication and the language disorder itself.

To gain insight into aphasia, empathy, and the type of interaction design that may best support such a tool, we conducted an exploratory study (guided demonstrations and interviews) with ten individuals (students, professionals, and researchers) with experience in aphasia (speech and hearing science/psychology), averaging 10.3 years of experience in aphasia/language disorders, with mean age of 34.6 years.

The initial prototype (Figure 13.1 & 13.2) was built in Java and simulated an IM conversation between the user (whose text was distorted) and Eliza (Weizenbaum, 1967), a simple and unbiased text-based computerized conversational partner. The system provided a simplistic approximation of the effects of aphasia by randomly dropping the user's function words (16-33%), randomly dropping any words (20-40%), and randomly replacing non-function words with other random words from a dictionary file. While this prototype was a simplified representation of some effects of aphasia, it illustrated the concept and functionality of ACES in order to gather informed feedback. Users saw how their messages were distorted through an IM window (Figure 13.1), and through a log of distortions (Figure 13.2).

Each participant met with a researcher for one session of approximately 40 minutes duration. Sessions began with an explanation/demonstration of the prototype. Participants then used the prototype to have a conversation with Eliza. Following the IM conversation, researchers had a discussion with each participant, and concluded with a brief questionnaire using a 7-point Likert scale (7 agree, 1 disagree). There was no remuneration for participation.

13.2.1 Results and Lessons Learned

Overall, participants agreed that this tool could be used to teach empathy and understanding for aphasic individuals to: friends (mean=6.1), family (mean=6.2), clinicians (mean=5.9) and professionals (mean=6.1). In addition, participants provided explicit guidance on system improvements and general suggestions on the project. For example, one participant stated that:

> *This could be shown to people outside of the community of communication experts and their clients. It could teach empathy and acceptance to a group/community as a whole. -P4*

Another participant stated that this software was critical for training clinicians:

> *I don't think treatment will be successful at all if the clinician is not understanding! They are the individuals who should be, and without that empathy, few gains will be achieved . -P5*

Another participant echoed P5 by stating that:

> *I believe that as a simulation of a disorder... this software could be used in classrooms as early as undergrad . -P3*

All participants stated the need for a customizable future system. Given the spectrum of aphasic disorders and the manifestation variety within each sub-disorder, such a tool needs to be robust enough to emulate different types of aphasia. Participants also felt that while the emulation would not need to be 'spot on,' it would need to emulate the key distortions of each condition reasonably well.

Participants suggested applications for an aphasia emulation system. One professional speech language pathologist suggested political applications, such as helping to raise awareness to increase both funding and accessibility. Another felt there to be a need for greater empathy by physicians and nurses in hospital and ER settings, who can seem dismissive of patient struggles. Most felt that an emulation system could be used by therapists with families that have a member with aphasia. All saw direct application of the software in the classroom. Some suggested that students use the software in class. Others suggested use at home, while attempting to communicate with friends for one evening. Participants mentioned that current practice in speech and hearing classrooms is to "role play" the effects of the aphasia (similar to acting). They reported that this approach is often awkward and inaccurate, indicating that an emulation system would be a large improvement over the status quo.

Perhaps the most tangible benefit of our study was the direct resources our participants provided. Students, faculty and professionals each mentioned key aspects of aphasia that should be emulated, if this system were to be useful.

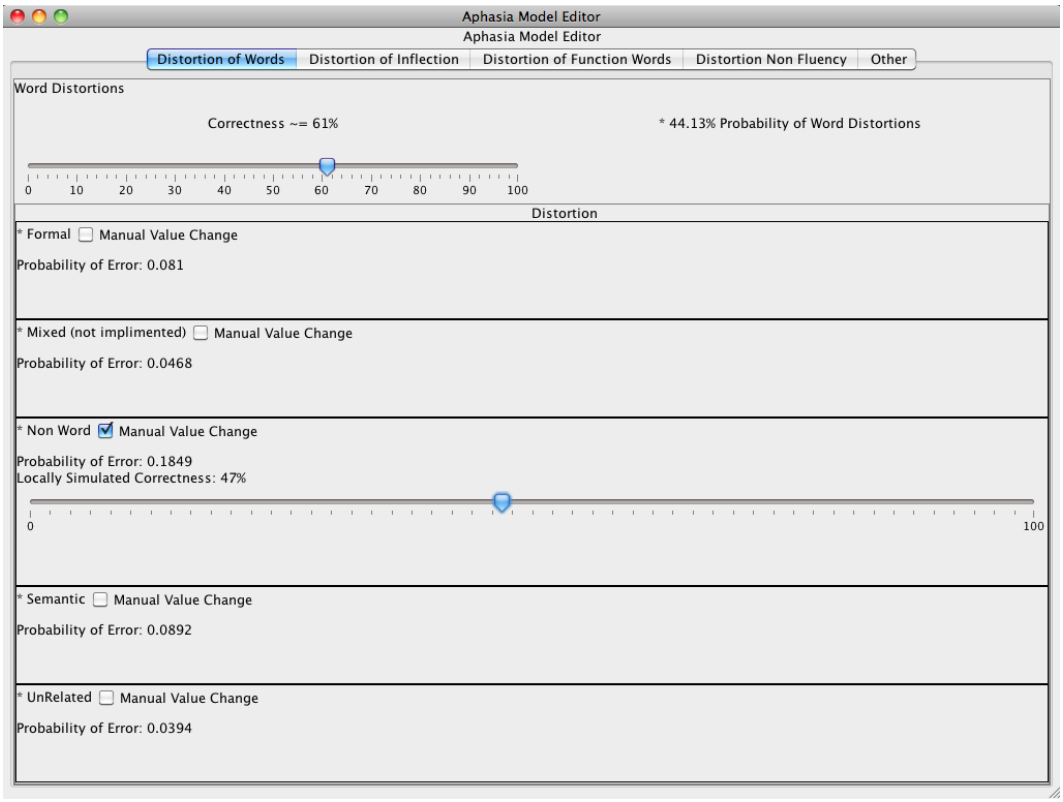


Figure 13.3: *ACES Model Editor - Distortion of Word Tab*
The overall correctness set to 61%, but the user has manually set the nonword errors to have a higher probability of occurring, increasing the overall probability of error to be 44.13%

13.3 ACES: Aphasia Emulation Software

ACES is a both an IM client and an instantiation of a model of linguistic distortions caused by aphasia. ACES distorts messages in a manner similar to those of an individual with aphasia. Our model determines the specific nature and rate of distortions, based upon feedback from study participants and the large body of related literature. In the following sections we describe the underlying model of aphasia distortions and the user interface components of ACES. While ACES was modeled on aphasia, our system could, in principle, be used to emulate other linguistic disorders.

Please note that the contribution of our work is to promote empathy and understanding in people through experiencing the linguistic challenges of aphasic individuals. It is not to perfectly emulate aphasia. Moreover, this project focuses on expressive language distortion, and does not address those with receptive language distortion.

13.3.1 Modeling The Effects of Aphasia

The effects of aphasia can vary widely based on the severity of the impairment, the type of aphasia, and even on the type of word that is subject to error. Therefore, we constructed a modeling system, **A**phasia **C**haracteristics **E**mulation **M**odel **E**ditor (**ACE-ME**), providing controls that allows the user to configure the degree and type of distortions that will be applied to their messages (Figure 13.3).

ACE-ME sub-divides the distortions into 5 conceptual categories: Distortion of Content Words, Distortion of Inflections, Distortion of Function Words, Distortion of Fluency, and Other. Each category defines error types that affect similar types of words (e.g., only function words or only content words). By grouping similar errors together, we leverage existing literature that focuses on specific distortion types and provide an intuitive interface for users. ACE-ME includes the breadth of distortions experienced by individuals with aphasia. Although not every aphasic linguistic error is included, our goal was not to create a linguistically exhaustive emulation tool, but one that was communicatively disruptive in a manner similar to those with aphasia. As knowledge of aphasia and natural language processing expand, so can the capabilities of our system.

Further, rather than creating a cognitive model of Aphasia itself, our goal with ACE-ME was to emulate individual distortions themselves[1]. Our theory was that a user experiencing the net result of the multiple types of distortions, the cumulative effect would appear "Aphasia-like." In many ways, our goal was to convince someone on the other end that they are having the promised experience, or talking with someone with aphasia. If we can active that experience, then *how* is less of a concern (be it cognitive modeling or emulating specific distortions).

13.3.1.1 Distortion of Content Words

Most forms of aphasia affect the production of content words (verbs, nouns, adjectives, and adverbs). However, the effect and frequency of each error type differs depending upon the disorder and severity. Schwartz (Schwartz et al., 2006) highlights five key ways that content word production can be distorted: Formal Errors, Nonword Errors, Semantic Errors, Unrelated Errors and Mixed Errors. ACE-ME implements the first four error types. We omit Mixed Errors due to their technical complexity and low incident rate.

- **Formal errors** occur when a user mistakenly says a content word that is phonetically similar, yet semantically different (Schwartz et al., 2006). For example: intending to say 'population' but

[1] However, in many ways the probabilities that went into the ACES and ACE-ME came from the cognitive model research done in prior work. In some "one off" way, we are using cognitive models.

stating 'pollution.' Formal errors are common with most aphasic individuals. ACE-ME emulates formal errors by utilizing a spell check system, JaSpell (Martins, 2010). We force JaSpell to return an alternative for a word users enter (even though spelled correctly). We then randomly select one of the possible alternatives as a replacement. By using a spell check system, the word replaced would be similar to the original word given that suggested "alternatives" are based on the spelling of the original word.

- **Nonword errors** occur when a generated word is not a valid English word (Schwartz et al., 2006). The generated nonword can be phonetically similar to the source word. For example: the word 'castle' is changed to 'kaksel.' Nonword errors are common for many aphasic speakers, particularly those with conduction or Wernike's aphasia. ACE-ME emulates a nonword error by randomly replacing a random number of letters in the original word. Vowels are replaced by other vowels, and consonants are replaced by consonant digraphs (e.g., 'sh,' 'ch' etc) or other consonants (excluding x and z, which are uncommon in English). All generated words are then verified to not be 'real words.'
- **Semantic errors** occur when the target word is replaced by another word which is semantically similar (Schwartz et al., 2006). For example: 'birthday' is replaced with 'anniversary,' or 'cake' with 'bread.' While anniversary is not a synonym for birthday, nor bread for cake, they are associated. This error occurs with all aphasic types. Because semantic errors are broader than synonyms, using a thesaurus would not capture the nuance of semantic errors. Therefore, ACE-ME identifies the root of the original word (Porter, 2001), searches the root in a ConceptNet database (Liu et al., 2010), and chooses a random word from possible semantic matches. ConceptNet is a NLP project that, among other things, attempts to group words together based on their semantic similarity. We extracted a relational database linking words to lists of semantically related terms.
- **Unrelated errors** are valid English words, that have no semantic or strong phonetic relationship to the original word (Schwartz et al., 2006). These errors occur particularly in severe cases of aphasia. ACE-ME randomly selects another content word from a list and replaces the original word with the unrelated term.
- **Mixed errors** are similar to formal errors, except the target words are semantically similar (Schwartz et al., 2006). For example: 'start' is distorted to 'stop,' or 'snail' to 'snake.' Mixed errors are are conceptually similar but programmatically very difficult. While we can readily generate semantically or phonetically similar words, we do not posses a large enough data set of either semantically or phonetically similar words to have a suitable intersection of the two. These errors are not very common, so not generating them should have little impact.

To understand how severity of aphasia correlates with frequency of word errors, Schwartz (Schwartz et al., 2006) performed an analysis of errors made in picture naming tasks. Aphasic subjects were shown a picture and asked to identify it. Schwartz examined the errors and created regression models to predict how likely a given error was to occur based on a subject's overall level of impairment. Using the raw data published in Schwartz's paper, we re-generated the probability of each of the five error types based on subject impairment. We used these probabilistic models in ACE-ME to calculate estimated frequency of each error. The user can move a "correctness" slider to set the probability of each of the five content word errors based on Schwartz's work. ACE-ME also allows the user to manually set each specific content word distortion (Figure 13.3). ACE-ME displays the total probability of any content word error, calculated by summing the probabilities of all possible content word errors. If the user has not manually adjusted an error rate, the probability of content word error is one minus the correctness percentage.

13.3.1.2 Distortion of Inflections

Grammatical inflection is the modification of a word to express different grammatical categories such as tense, mood, voice, aspect, person, number, gender and case. English inflections are usually suffixes, such as the plural inflection on nouns ("apple" versus "apples").

ACE-ME distorts only verb inflection, the most common type of inflection error for aphasic individuals (especially for those with agrammatic speech). An example of such a distortion is changing "running" to its base form, "run." ACE-ME emulates verb inflection distortion by stemming verbs (Porter, 2001) and using the verb's infinitive form. Users can set the probability of Inflection Distortion with an interface slider labeled "Inflection Morphology of Verbs."

13.3.1.3 Distortion of Function Words

A common error made by aphasic individuals (especially those with agrammatic aphasia) is to omit function words. Function words are pronouns, articles, conjunctions, auxiliary verbs, interjections, particles, expletives, and prepositions. ACE-ME removes randomly selected function words. Users can set the probability of Function Word Distortion with an interface slider labeled "Dropping Function Words."

13.3.1.4 Distortion of Fluency

The fluency of sentence construction can also be affected by aphasia. ACE-ME targets three common types of distortions that can affect sentence fluency: Conduit d'approche, Omissions, and Semantic Description.

- **Conduit d'approche:** This commonly affects those with conduction aphasia. Conduit d'approche occurs when an individual adds/removes or distorts the production of a word, then iteratively repairs the errors until arriving at the intended production (Goodglass et al., 2001). For example: a series of Conduit d'approche would be "brepple," "gresles," "glasles," and then completing the effort with "glasses." ACE-ME emulates Conduit d'approche errors by repeatedly applying a random number of nonword errors to the same word. The errors are then replayed in reverse order in the sent message, giving the appearance of correcting the specific word. Each attempt is put in a separate message followed by ellipses. The slight pause gives the impression that the sender is attempting to correct the word. Once the word is "correct," the remainder of the message is sent.
- **Omissions:** Most aphasic individuals omit or drop words from their sentences (Sarno, 1998). To emulate the effects of random omission, ACE-ME randomly drops words from a user's message.
- **Semantic Description:** Many aphasic individuals can describe the features of a target word when they cannot recall the specific word (Sarno, 1998). This is a semantic description error (unlike a semantic error which is a one-to-one substitution). For example, an individual trying to describe glasses may say, "seeing, head, face, eyes, round." ACE-ME utilizes the ConceptNet database (Liu et al., 2010), to search for a random number of semantically related terms and replaces the original word with a set of semantic descriptors (separated by ellipses).

Users can set the probability of all errors related to Fluency using ACE-ME interface sliders. As there is more than one type of Fluency error, ACE-ME also displays the total probability of a Fluency error occurring.

13.3.1.5 Other

There are two other effects common to aphasia that ACE-ME emulates. These are pauses and distortion awareness. Aphasic individuals often pause at atypical times. ACE-ME emulates these interruptions by randomly breaking up a message, sending each part as a separate IM.

People with certain types of aphasia (e.g., Wernicke's) can be unaware of the errors in their speech, while others (e.g., Broca's) are generally aware of their errors. To allow users to experience the frustration

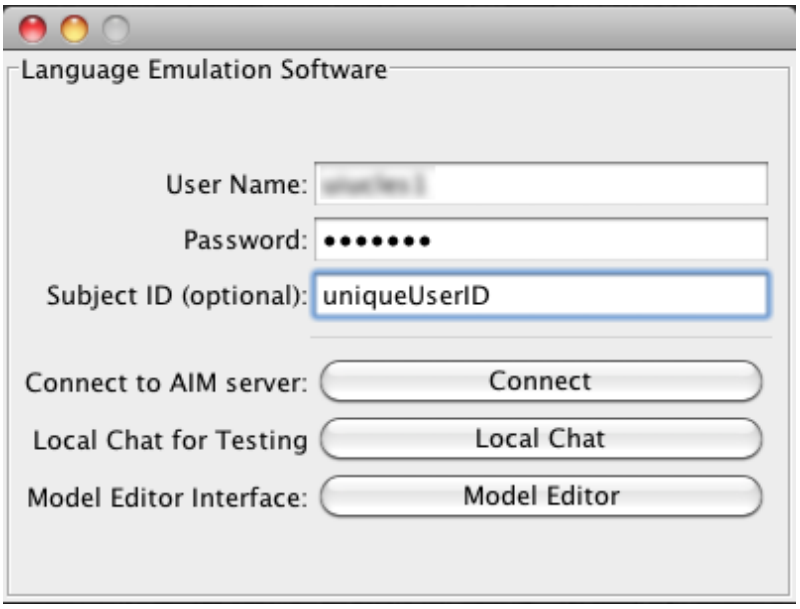


Figure 13.4: *ACES Admin Window*

of not knowing why their conversation partner is confused (by not seeing the distorted message), as well as the frustration of not being able to send the intended message (e.g., seeing the distorted text), ACE-ME allows the user to toggle between displaying the original or the distorted message.

Users can set the probability of pauses by using the interface slider. Distortion awareness can be toggled by an interface check box in ACE-ME.

13.3.2 ACES Interface Components

Our system has three main components: an IM Window for engaging in instant message conversations; a Model Editor for configuring the distortions; and an Admin Window for login and switching between the IM and the Model Editor.

Upon launch, users are presented with an Admin Window (Figure 13.4). Users can input an AOL Instant Message user name, password, and a subject ID (used for logging all conversations to protect privacy). Once login information is entered, users can launch the other two components.

The ACES Model Editor (Figure 13.3) consists of five tabs corresponding to the five categories of distortions. Users can switch between tabs and interact with all sliders and check-boxes. All changes made to the model editor affect the current IM conversation in real-time.

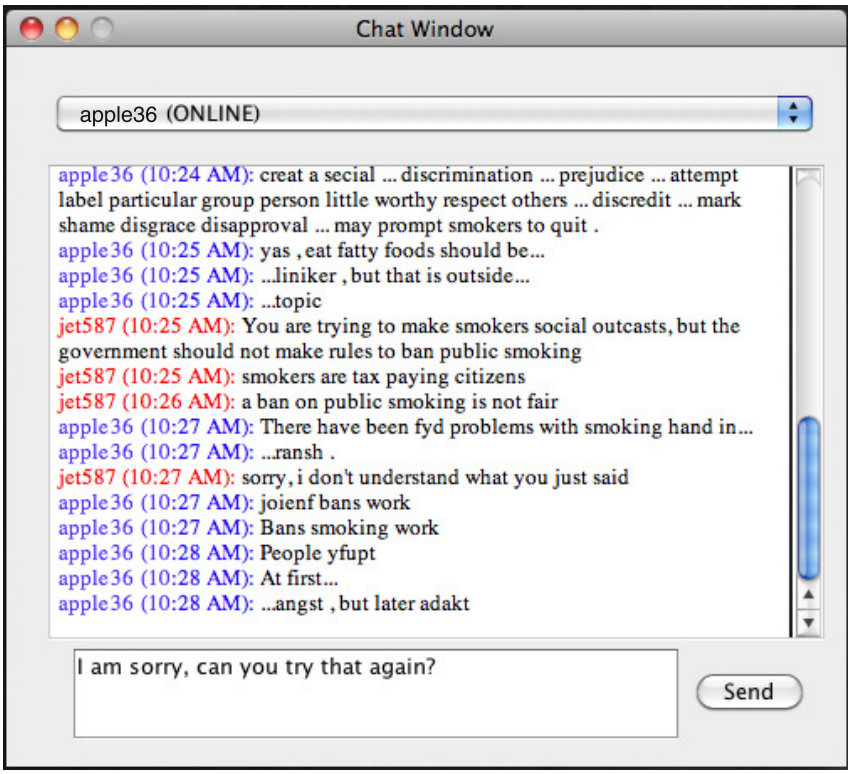


Figure 13.5: *ACES Instant Message Window*
The current (red) user's partner (blue) has their text distorted.
Note: this is a log from an actual experiment conversation (See Chapter 14)

The IM Window (Figure 13.5) is a standard IM interface with a message history, text input field, send button, and buddy list. IM history is color coded for easy identification of message origin (red=user, blue=conversation partner).

13.3.3 Flexible Implementation

Our system leverages a plug-in architecture for distortions. Researchers, instructors, or therapists can create or customize a specific distortion type. The ACES system examines a directory where the distortion descriptions reside, including any new distortion from this directory.

ACES sends IMs over the AOL Instant Message network, though it can easily be extended to other protocols, because it leverages the JBuddy library (Software, 2010) to facilitate connection with the AOL servers. JBuddy also supports ICQ, MSN, Yahoo, Google Talk, XMPP (Jabber), Lotus Sametime, Microsoft OCS 2007, and LCS 2005.

ACES was implemented using Java 1.6, allowing the software to execute on nearly any machine, though our system has only been tested on Apple's OSX. As a result, teachers, medical staff, and researchers need not invest in new hardware to run ACES.

13.3.3.1 ACES Logging

ACES is also conducive to post-conversation analysis and reflection because it logs all conversations in an HTML file of the perceived conversation (the user's sent messages), and in an XML file of the conversation (including the original message, the distorted message, and which distortion were applied). The HTML version is ideal for analyzing the conversation and reflecting upon difficulties in the user experience. The XML file allows researchers to analyze how users changed their behavior while trying to communicate with aphasic distortions.

Chapter 14

Promoting Empathy Towards Aphasia

In order to observe the effects of using ACES on awareness and empathy, we conducted an in-depth user study. Sixty-four individuals (grouped in subject-pairs) engaged in IM conversations with each other. We wanted to see whether experiencing text distortions first-hand through IM conversations had an impact on a subject's level of empathy for individuals with aphasia. To test this, we utilized a between-subject 2x2 factorial design. The first factor in our factorial design compared a **Treatment Group** (where subjects experienced distortions first-hand) to a **Control Group** (where subjects did not experience aphasic distortions). The second factor in our factorial design compared subjects who had prior knowledge of aphasia, through formal education or personal exposure (**Informed Sub-population**) with subjects who did not (**Uninformed Sub-population**). Our *Independent Variable* was whether or not participants experienced distortions by ACES. All participants (both treatment and control groups and both informed and uninformed sub-populations) experienced the same experimental protocol, as they were given the same prompts. Therefore subjects were blind with respect to which group (treatment or control) they were assigned. Both subjects in any given subject-pair were from the same sub-population (Informed/Uninformed) and same group (Treatment/Control).

Some research, content, and text from this Chapter is reproduced from (Hailpern et al., 2011a)

	Description
Step #1	Demographic Questions
	Empathy Evaluation
	Pre-Study Questions on Aphasia
Step #2	Conversation with Partner, Prompt #1
Step #3	Conversation with Partner, Prompt #2
	Participant Roles are Switched
Step #4	Final Set of Questions

Table 14.1: *Study Session Order*

14.1 Experimental Protocol & Design

Each study session consisted of four steps lasting a total of 45-60 minutes (Table 14.1). A session involved IMs between two participants (a subject-pair) who did not know each other. Subject-pairs remained physically separated. Each member of a pair was placed in separate identical rooms with a 23" iMac computer. All questionnaires were digital and administered using third-party software.

All participants completed both a demographic survey and a pre-study questionnaire to assess empathy using Mehrabian's measure[1] of emotional empathy (Mehrabian and Epstein, 1972). Each participant also answered questions about their knowledge and background on aphasia. This allowed us to ensure equal background, pre-knowledge, and general empathy level across treatment and control groups. Pre-study demographic and background measures were not utilized to assess ACES effect on awareness and empathy.

Upon completion of the pre-study questionnaires, all participants were given an identical explanation of the study protocol. All participants (regardless of group or sub-population) were told that they would have two ten-minute IM conversations with each other [2]. During these conversations one member of the subject-pair would take on the **Aphasia Role**, while the other member of the subject-pair would take on the **Typical Role**. After the first conversation, they would switch roles (ensuring that each participant would play both roles). All subjects (in both control and treatment groups) were told that the participant in the Aphasia role would have his/her text distorted by ACES, as though they had aphasia.

To guide their conversations, participants were instructed to have a "debate" or "discussion" around a specific topic. Further, each participant was assigned a position (**Pro** or **Con**), and was provided a suggested list of talking points to support their position (participants could also use their own talking points). Subjects were assigned one topic for the first conversation, and another topic for the second. The two debate topics were:

- *The age at which people gain the right to vote should be lowered to 16 years*
- *Smoking should be banned in all public places*

To conclude the study explanation, all participants were told *"Remember aphasia can be mild to severe, creating distortions that are obvious, to those which are not so easily noticed or apparent."* This statement

[1] Mehrabian's quantitative measure of empathy assigns an individual an empathy score based on responses to 33 questions. Scores range from -132 to +132. According to Mehrabian, a representative population should yield a mean score of 33 and S.D. of 24.

[2] Participants were logged into IM accounts created for this study, not requiring subjects to disclose their own user names or passwords.

was included in the instructions so participants in the "control" group would not be concerned if they did not notice any obvious distortions.

At the end of both IM conversations, participants were administered a single questionnaire to gauge their reaction to ACES and aphasia. Participants were then remunerated with a $10 gift certificate to Amazon.com.

14.1.1 Dependent Measure

A single post-study questionnaire was utilized to assess the effects of using ACES on awareness and empathy. Ten questions on the post-study questionnaire were scored on a 7-point Likert scale. Additionally short-response questions were included to obtain qualitative reactions to ACES.

14.1.2 Factorial Design, Counterbalancing, Treatment Effects

Our cohort of sixty-four individuals consisted of two sub-populations. Thirty-two subjects (one half of our subjects) had prior knowledge of aphasia either through formal education or personal exposure (Informed Sub-population), and the other thirty-two subjects did not (Uninformed Sub-population). Further, half of each sub-population was in the Control Group, and half was in the Treatment Group. Thus, we used a **2x2 Between-Subject Factorial Design** with 16 subjects (8 subject-pairs) in each of the following **Experimental Conditions:** Informed-Treatment, Informed-Control, Uninformed-Treatment and Uninformed-Control. By testing both a Informed and Uninformed population, we had the ability to test the potential applicability of ACES to both family members and clinicians/doctors in training.

Given the number of other factors inherently present in this type of experimental setup, there were at least three potential confounding effects (debate topic order, debate position and role order). To help control for these potential confounding effects, we used counterbalancing, a method commonly employed to help avoid confounding from order of task and presentation. When counterbalancing, all permutations of the confounders are included in an attempt to minimize any bias due to these confounders that are not central to the experimental question. Thus, for each Experimental Condition (e.g., Informed/Uninformed and Treatment/Control), we needed 8 pairs of conversations to account for all permutations of Role (Aphasic/Typical), Topic (Voting/Smoking), and Debate Position (Pro/Con).

In addition, we attempted to control for "treatment effects." A treatment effect occurs when a participant has a reaction to a placebo simply because they are told they are going to receive a treatment. By

effectively telling both Treatment and Control groups that they were going to be in the "treatment group," we attempted to control for the treatment effect. If we noticed a positive or negative response from participants in our control group, we would have confidence that it would be due to treatment effects, thus allowing us to better qualify the effects reported in the treatment group. Likewise, if no effects were observed in the control group, we would feel confident that the simple act of using IM in this experiment did not contribute to those results, and were due directly to ACES emulating the effects of aphasia.

14.1.3 Statistical Tests

To examine the effects of ACES we compared responses from the four Experimental Conditions to the post-study questionnaire in Step 4 (Table 14.1). A 2-way ANOVA (analysis of variance) was used to to test for differences between the four Experimental Conditions (Group by Sub-Population). An ANOVA test assumes that data are normally distributed and is robust to deviations in normality. Data is presented as median, inter-quartile range (the range of values that span the middle 50% of the data).

14.1.4 Subjects

Sixty-four participants completed the ACES experiment (Table 14.2). More than 75% of the Informed population had an educational background in psychology, speech and hearing science, or traditional sciences. Participants in the Uninformed population came from a wide variety of areas spanning engineering, science, and liberal arts.

The empathy scores of our subject population, as measured at the beginning of each session (Table 14.2), closely follow those of a standard population as presented in the original paper by Mehrabian (Mehrabian and Epstein, 1972). With a balanced male/female population, Mehrabian expects a mean score of 33 (standard deviation of 24). Our study cohort had a mean empathy score of 34.5 (standard deviation of 25.8). This is nearly identical to Mehrabian. To ensure control and treatment groups were equivalent, we examined the empathy scores of the groups and the four initial self-response questions regarding participant knowledge of and sensitivity to the challenges of communication with aphasia. We examined any differences overall and within the Informed/Uninformed demographics. There were no statistically significant differences ($p = 0.08$, $z=1.75$) between the control and treatment groups in a priori disposition, knowledge of, or empathy towards individuals with aphasia (results not shown).

	% Male	Age (SD)	Empathy (SD)
Overall	39.06	26.05 (7.30)	34.52 (25.82)
Overall - C	28.12	25.06 (7.27)	40.09 (22.60)
Overall - T	50.00	27.03 (7.30)	28.94 (27.92)
Informed	37.50	25.35 (6.83)	39.25 (25.38)
Informed - C	25.00	24.13 (6.54)	46.81 (21.28)
Informed - T	50.00	26.56 (7.11)	31.69 (27.51)
Uninformed	40.62	26.75 (7.78)	29.78 (25.76)
Uninformed - C	31.25	26.00 (8.04)	33.38 (22.50)
Uninformed - T	50.00	27.50 (7.69)	26.19 (28.95)

	Highest Level of Education Completed (%)		
	High School	**Bachelors**	**Graduate**
Overall	25.00	56.25	18.75
Overall - C	28.12	56.25	15.63
Overall - T	21.88	56.25	21.87
Informed	34.38	53.12	12.50
Informed - C	43.75	37.50	18.75
Informed - T	25.00	68.75	6.25
Uninformed	15.62	59.38	25.00
Uninformed - C	12.50	75.00	12.50
Uninformed - T	18.75	43.75	37.50

Table 14.2: *Population Demographics & Empathy Scores*
Empathy refers to Mehrabian's measure of emotional empathy (See Section 14.1).
Mean age and empathy scores are shown.
C = Control T = Treatment

		Control Group			Treatment Group			2-Way ANOVA	
		*M*edian		Mean (SD)	*M*edian		Mean (SD)	F Statistic	p -val
1	In hindsight, this experience made me < > empathetic to individuals with aphasia.†	5	[4,5]	4.75 (0.80)	6	[6,6]	5.91 (0.77)	F(1,61)=34.84	0.0001
2	In hindsight, this experience made me < > empathetic to people who try to understand an individual with aphasia. †	4	[4,5]	4.62 (0.87)	6	[5,6]	5.59 (0.87)	F(1,61)=19.72	0.0001
3	This tool is useful for building empathy for family members‡	4	[4,5]	4.19 (1.47)	6.5*	[5,7]	6.09 (1.06)	F(1,61)=34.96	0.0001
4	This tool is useful for building empathy for speech therapists‡	4	[4,5]	4.47 (1.59)	6	[5,7]	6.09 (1.03)	F(1,61)=23.28	0.0001
5	This tool is useful for building empathy for nurses and doctors‡	4.5*	[4,5]	4.47 (1.54)	7	[5,7]	6.22 (0.91)	F(1,61)=30.13	0.0001
6	This tool is useful for building empathy for care givers‡	4	[4,5]	4.38 (1.50)	7	[6,7]	6.38 (0.91)	F(1,61)=41.12	0.0001
7	This tool is useful for education/training of individuals who treat/interact with individuals with aphasia‡	4	[4,6]	4.5 (1.65)	7	[6,7]	6.53 (0.80)	F(1,61)=39.02	0.0001
8	Which role helped you understand aphasia better?§	4	[3.5,4]	3.91 (1.42)	2	[1,3]	2.28 (1.55)	F(1,61)=18.79	0.0001
9	Which role increased your empathy for the the perspective of understanding an individual with aphasia? §	4	[4,4.5]*	4.5 (1.03)	1.5*	[3.5,6]*	3.72 (2.40)	F(1,61)=1.02	0.3169
10	Which role increased your empathy for the the perspective of an individual with aphasia? §	4	[3.5,4]*	3.69 (1.38)	2	[1,4.5]	2.72 (2.25)	F(1,61)=4.26	0.0433

Table 14.3: *Empathy Study Results*

Median represents median value and interquartile range [IQ range] - 2-Way ANOVA Values represents Treatment vs. Control Factor

2-Way ANOVA Values for Informed Sub-population vs. Uninformed Sub-population (not shown) all p>0.05

* *Half values are due to even number of data points where the vale separating the higher half from the lower half lies between two different value, resulting in a median which is the average of the two values. For example a data set of [1,4,4,5,5,7] would have a median of 4.5 although 4.5 is not a possible value.*

†Question Choices Were: 1=Much Less, 2=Less, 3=Slightly Less 4= No Change, 5= Slightly More, 6=More, 7=Much More

‡Question Choices Were: 1= Disagree, 4=Neutral, 7 = Agree §Question Choices Were: 1= Aphasia, 4=Equal, 7=No Distortions/Normal

14.2 Empathy Study Results

Overall, participants in the group experiencing aphasic distortions by the ACES were strongly affected by their experience relative to the control group (Table 14.3, Questions 1 & 2). For example, participants indicated that the experience made them more empathetic to individuals with aphasia compared to the control group who felt only a slight change, with $p<0.0001$. Participants in the treatment group strongly agreed that ACES should be used in all applications explicitly listed, while the control group felt neutral as to the applicability of ACES (Table 14.3, Questions 3-7). Participants in the Treatment group strongly agreed that ACES could be used to increase empathy in caregivers, while participants that did not experience distortions felt neutral towards ACES use, with $p<0.0001$. Participants in the Treatment group felt that they gained understanding and empathy from having their text distorted, while the control group did not rank one role higher than another (Table 14.3, Questions 8 & 10).

When data was analyzed separately for the Informed and Uninformed groups, results were similar to those of the full cohort (results not shown). Finally, we compared the Informed treatment group with the Uninformed treatment group to see if a priori knowledge of aphasia influenced subject response. Both treatment groups found ACES beneficial in increasing knowledge and empathy.

Though not shown in Table 14.3, there was no significant difference in any of the post-study questions by Informed/Uninformed Sub-Population. Furthermore, there were no interaction effects between the Treatment Group (Treatment/Control) and Sub-Population (Informed/Uninformed) (data not shown - available from the authors upon request).

Figure 13.5 is a representative conversation between two subjects (taken from our study). The IM window is from jet587, whose partner, apple36, is having his text distorted.

14.2.1 Post Hoc Results

The one question with a non-significant value overall was question 9: *"Which role increased your empathy for the the perspective of understanding an individual with aphasia?"* Given the large inter-quartile range within the Treatment group, we hypothesized that the lack of significance was due to learning effects (having more exposure to the intervention makes you increasingly more empathetic). To test this hypothesis, we subdivided the Treatment Group by comparing the responses of subjects who were in the Aphasic Role first, to those in the Typical Role first. Because these are mutually exclusive subsets of the Treatment group we performed a between subject analysis comparing responses with a

unpaired student T-test and found p=0.015 (-2.66=t). In particular, the mean response (2.69 + 1.99) of participants whose first tole was Typical, indicated that the Aphasic role increased empathy more. In contrast, the mean response (4.75 + 2.38) of participants whose first tole was Aphasic, indicated that the Typical role increased empathy more.

Thus, whichever role the participant experienced second was the role reported to have increased their empathy more. This post-hoc analysis suggests that it does not matter which role (Aphasic or Typical) participants take on first. Rather, we hypothesize that it is the opportunity to play both roles that may increase empathy. This would need to be validated in future experiments. We tested the other questions for learning effects and did not find any.

14.3 Empathy Study Discussion

Participants who experienced the distortions of aphasia had a much stronger response than those in the control group. Those participants in the treatment group reported "strong" effects from the experiment, including increased empathy and envisioning a wide variety of potential uses of ACES. In contrast, participants who did not have their text distorted reported little or no change in their perception of or empathy towards aphasia. The median quantitative responses from the control group were almost unanimously at 4 (a neutral opinion/no change), with little deviation in inter-quartile ranges. The control group also saw no application for ACES to increase empathy or awareness of aphasia. This is in stark contrast to those with exposure to aphasic distortions.

There is always the possibility of a change in empathy resulting from being in an experimental condition (known as experimental demand). However, given the great difference in responses between the two groups (qualitatively and statistically), we are confident that ACES demonstrates a real impact on increasing awareness and empathy by experiencing distortions caused by aphasia. This is further supported in that both groups were given identical prompts, and the experiment was fully counterbalanced.

While quantitative feedback provides strong evidence of the effectiveness of ACES, qualitative feedback highlights how informative the ACES experience was for participants, and how supportive they are of the project. For example, one participant from the Informed Treatment group stated:

> *This was a wonderful way to gain empathy! I've learned about aphasia in classes, but this perspective was very helpful and will make a lasting impression.*

The qualitative feedback supported how prior experience with aphasia was not a prerequisite for gaining insight. One participant from the Uninformed Treatment group stated:

> *It was the most eye-opening from the point of view of having my text distorted. It was amazing how hard it can be sometimes to get a point across. From the aphasic side it is almost like experiencing it from both sides, because you see how difficult what you say can be for the 'normal' person.*

Participants also suggested that even short exposure could make a lasting impact, saying, *"it would only take a few minutes to gain perspective."*

In contrast, qualitative feedback from subjects in the control group supported our quantitative findings, indicating that this was not a meaningful experience. Given the stark difference in qualitative and quantitative responses between the control and treatment groups, we believe that ACES can provide a meaningful experience for all users who seek to better understand and empathize with aphasic individuals.

14.4 Future Work

Our experiment examined the impact of ACES on participants' empathy and understanding of aphasic individuals. However, our long term goal is to meaningfully improve real-world interactions, thereby improving quality of life and therapy for individuals with aphasia. We propose conducting a follow-up study to explore whether a subject's experience with ACES can positively impact their real-world interactions with individuals with aphasia. This study would directly involve individuals with aphasia in order to determine if they can perceive an increase in empathy present in users of ACES. We further wish to investigate the long-term impact of ACES, and how to integrate the software within a classroom of therapeutic setting

While ACES supports a robust set of distortions, there are always refinements and less common errors that can be implemented. Over time, we hope to increase the capabilities and functionality that ACES provides.

Though this project explicitly targets aphasia, our goal is to introduce language distortion emulation as a new approach to increasing empathy for those with other language impairments. We envision, for instance, a system similar to ACES that distorts text as though it comes from a young child. This can teach patience, understanding, and empathy to parents, teachers, and caregivers.

14.5 Conclusion

Empathy and understanding from family members, friends and clinicians can enhance the quality of life of aphasic individuals. Family members can deny or underestimate the severity and presence of aphasic errors. Without empathy, the quality of speech therapy can suffer, jeopardizing the speed and recovery of aphasic individuals. Our work has made several contributions to address these concerns stemming from a lack of empathy.

First, we leveraged speech-language and psychological theory to design and construct a model of aphasic distortions. Second, from an initial investigation we refined our model and designed a system to increase awareness and empathy with aphasic individuals. Third, we developed a novel system (ACES) that allows a neurologically typical individual to experience firsthand, the linguistic distortions of aphasia. Fourth, ACES was validated in an investigation with 64 participants (with and without background on aphasia). Results from this study strongly show that using ACES can increase empathy and awareness of aphasia.

Our model and system demonstrate how technology can play a central role in increasing empathy, awareness and understanding for individuals with a language disorder. Our approach can be used in many other domains where atypical language is present and can be emulated. It is through empathy that we learn to understand each other.

Chapter 15

Aphasia Emulation, Realism, and the Turing Test

While the ACES empathy research in Chapter 14 presented the first language disorder emulation system and its impact on empathy, it did not validate the quality or realism of the distortions ACES applied. This chapter seeks to demonstrate how discernible ACES distortions are from actual statements generated by individuals with aphasia. If we demonstrate that distortions users experienced were realistic, we will increase the impact and validity of our original study. Further, as ACES appears to be nearly indistinguishable from realistic distortions, it indicates that ACES may prove a valuable and realistic aid for increasing empathy for family members, friends, clinicians in training and other caregivers.

This chapter illustrates the "realism" of ACES distortions in two ways. First, we perform a Turing Test study in which participants must distinguish samples of distorted text generated by a human from samples of text distorted by a computer. Much like the original Turing Test proposed by Alan Turing (Turing, 1950), if participants cannot reliably tell the origin of distortions (whether computer or human generated), the computer could be said to have passed the test. Second, we ask participants to explicitly rate the realism of distortions in text samples on a Likert scale. If participants rate both computer and human generated text samples as being equally realistic, it adds further quantitative support to the realism of ACES distortions. The foremost contribution of this chapter is the demonstration that ACES generates realistic aphasic distortions, thus validating the applicability of ACES as aphasic emulation software and supporting the feasibility of other language emulation software research.

15.1 Research Question & Motivation

The initial ACES user study examined the impact of ACES on empathy and awareness of aphasia. However, there was no examination of the distortions themselves, their perceived quality, or their

Some research, content, and text from this Chapter is reproduced from (Hailpern et al., 2011a)

realism. While ACES' distortions were deeply grounded in literature (providing both the probabilistic underpinnings and the manner in which text is distorted), creating a novel and unique form of language emulation has the potential to produce distortions at varying degrees of realism. Therefore this chapter seeks to answer the following questions:

Can users differentiate computer-generated distortions from distortions generated by individuals with aphasia?

How realistic are the distortions of aphasia generated by ACES?

We answer these questions with a two-step experimental design, detailed in the following section. We then describe our target population, outline the methods for analysis, and present results for each of our two experiments. Discussion of our results follow, along with implications for future work.

15.2 Experimental Design

To answer our research questions, we recruited 24 participants to examine distortions generated by ACES, utilizing an online questionnaire. Each questionnaire presented users with a set of demographic questions, in addition to 48 data generating questions. Each page of the online questionnaire contained only one data generating question. A data generating question consisted of a text sample and a question about the sample. One half of the text samples were generated by ACES, while the other half were taken from transcripts generated by individuals with aphasia. These two halves will be referred to as the **Computer Group** and the **Human Group**, respectively. In an experimental context, these can be thought of as a treatment group and a control group. If ACES distortions are indistinguishable from human generated distortions, which is the goal of this project, *the desired outcome is to see a lack of statical significance ($p \geq 0.05$)* when comparing the Computer Group to the Human Group.

To control for order effects, we presented the questions in one of two sequences. The question order within each sequence was created randomly. Users were randomly assigned to one of the two presentation sequences. No text sample was presented to a user more than once (to control for learning effects). At the end of the study, participants received a $7 Amazon.com gift certificate.

The next section describes the types of aphasia targeted in this study. We then detail the two sets of questions asked of participants, and discuss the origin of all text samples presented to users.

15.2.1 Types of Aphasia

Based on the Boston classification system, there are numerous types of aphasia that can be broadly categorized as non-fluent aphasia (Agrammatic, Broca's, Transcortical Motor, Global) or fluent aphasia (Wernicke, Transcortical Sensory, Conduction, Anomic). Individuals within each subtype of aphasia have distinctive characteristics to their speech, and the errors that are made. We therefore wished to control for aphasia type in this experiment. Rather than tackling all known subtypes, we focused on two of the more common **Types of Aphasia**: Agrammatic and Anomic aphasia. Individuals with Agrammatic aphasia generally have difficulty with sentence structure and proper grammar, while having no difficulty with word selection. Common errors include difficulties in verb tense, dropping function words, inconsistency with the length or fluidity of sentences, and incorporating many breaks and pauses. Individuals with Anomic aphasia have difficulty with selecting and producing correct content words, though their grammar is generally correct. For example, words may be replaced by other words that are semantically related ('birthday' with 'anniversary' or 'cake'), that have no semantic relationship (cat with airplane), that have similar phonetic sounds ('population' with 'pollution'), or non-words ('castle' with 'kaksel'). The availability of Anomic and Agrammatic aphasia transcripts, in addition to their general prevalence of individuals with these types of aphasia, influenced our selection.

15.2.2 Test Questions

The 48 data generating questions were evenly divided into two distinct **Tests**: *The Aphasia Turing Test* and the *"How Human" Test*. Each user first answered the 24 *Turing Test* questions, followed by 24 *"How Human"* questions. Each question was presented on a separate page. This limited participants ability to make judgments based on the other questions' distortions. Further, participants were unable to return to prior questions, thus preventing them from changing their answers. The questions and the presentation of each Test are detailed in the following sub-sections. At the end of each test, we asked users to describe their approach for answering the Tests' questions.

15.2.2.1 Aphasia Turing Test

For the *Aphasia Turing Test*, participants were presented with 24 text samples. For each text sample participants were instructed, *"for each sentence, please mark if it is 'Human' or 'Computer' in origin."* The following example is an actual text sample used in this test, with the distortion generated by ACES emulating an anomic individual with aphasia:

> *Well she was a kaur girl and she qobred in a house where she was mopping the under foot. Then there was something about a shoe and when she wore it she would be Cinderella. uh... She was told that a mumpkur would be her stagecoach... um... and little rats would be a horse. And so she went up to the castle. Her new shoes um... uh... fit uh... um...her like a... grove.*

Participants *were* told that some text was generated by an actual individual with aphasia (Human Group text samples), and some text was distorted by ACES (Computer Group text samples). Participants were not told it was a 50/50 split of text samples from the Human Group and Computer Group. Section 15.2.3 details how text samples were generated/collected.

In this regard, this experiment was designed as a variation of the Turing Test as proposed by Alan Turing in 1950 (Turing, 1950). The goal of this Test was to determine if our subjects could reliably differentiate machine from human. To "pass" the Turing Test, we would expect to see approximately a 50% accuracy at labeling text as computer or human (with no statistical difference in the accuracy between groups). Since modern computers cannot reliably pass the Turing Test, we did not hypothesize a priori that ACES would completely pass our Turing Test either. Even without passing the Turing Test, results can illuminate the believability of ACES' distortions, and if there is one type of aphasia (see Section 15.2.1) which ACES emulates more successfully.

15.2.2.2 "How Human" Test

For the *"How Human" Test*, participants were presented with 24 pairs of text samples (one pair per page). For each pair of text samples, the first was labeled as "Original Text" and the second as "Distorted Text." The "Original Text" was undistorted, while the "Distorted Text" had aphasic distortions applied to it. The following example is an actual text sample used in this test, with the distortion generated by ACES emulating an agrammatic individual with aphasia:

> **Original Text:** *Well the man is trying to wake up because of the alarm clock. And then he goes back to sleep. His wife is angry. Then the man eats breakfast, while his wife is showing him the time on the clock; the wife is saying "hurry up." And the man running out onto the street to get to work. The man was so tired, he goes to sleep at the office.*

> **Distorted Text:** *uh... the er... uh... is tri... wake up because the alarm clock uh... And... he goes back uh... His wife is angry, Then the man eat uh... breakfast, his wife is ah... eh... showing him the time... the wife is saying "um... up." uh... er... And the man running ah... onto the street to get to work. The uh... was so tired he goes to sleep at the office.*

Participants were asked to help researchers "improve" the distortion algorithms by rating how "human" the distortions appear on a Likert scale from 1-5 (where 1 is indistinguishable from a human who has aphasia, and 5 is unquestionably a computer). Like the *Aphasia Turing Test*, one half of the "Distorted Text" samples were generated by an individual with aphasia (Human Group text samples), and the other half of the text samples were distorted by ACES (Computer Group text samples). Section 15.2.3 details how both the original text and distorted text were generated/collected.

By design, this task forces users to make an implicit judgment call about the origin of each distorted text sample: "was this text distortion generated by a human or by a computer?" To allow participants to focus on the realism of the distortions/errors in the text samples rather than puzzling over their source, participants were told that all text was distorted by ACES. This deception allows us to objectively measure the realism of ACES distortions (Computer Group). It also provides an objective and comparable benchmark of human distortions (Human Group).

15.2.3 Text Samples

All text samples used in this experiment were extracted from published transcripts of individuals with aphasia (Boller and Grafman, 2001; Menn and Obler, 1990) or from the unpublished data files used in (Menn et al., 1998a; Menn et al., 1998b), which were provided to the researchers by Lise Menn, University of Colorado. Some transcripts were from picture describing tasks from the Wechsler Bellevue Intelligence Scale (Wechsler, 1955). Other text samples were from transcripts of individuals with aphasia reading children's stories[1]. The remainder of the text samples were from individuals with aphasia narrating children's stories from memory.

For each text sample, an original or intended version of the text was generated. If the transcript was from an individual reading a story, the read text was used as the original version. For transcripts that were not of an individual reading text, researchers attempted to fix the errors and create a non-distorted version of the same text (following similar sentence structure, word choice, and phrasing). For the "*How Human" Test*, these original versions of the text samples were used. The original versions of each

[1] Reading does produce errors in speech production (Menn and Obler, 1990).

text sample were also used as the basis for the ACES distortions. Each non-distorted, or error-free text sample would be sent through ACES, thus applying the ACES distortions to the text. Section 15.2.3.1 details the procedure for choosing and applying ACES distortions so as to ensure the distorted texts used were not “cherry picked." There were therefore three versions of each text sample, the aphasic version, the ‘original’ version, and the ACES version. Selection of text samples to be included in the experiment was random, thus not giving any preference to ‘more believable’ ACES text.

Our text samples came from six individuals. We ensured that there were an equal number of text samples from each individual within both Tests (e.g. aphasiac individual Alice contributed four text samples to the Turing Test, and four text samples to the “How Human" Test). Further, the text samples taken from each participant were split evenly across the Human Group and Computer Group (e.g. if aphasiac individual Bob contributed eight text samples, four were used directly from his transcript and four were used to construct an undistorted text sample, which was then distorted by ACES). No text sample was repeated across Tests or within a Test, and only one version (the true aphasic version or the computer version) of each text sample was used.

15.2.3.1 Generating ACES Distortions

For each set of transcripts generated by one individual, researchers constructed an ACES model that attempted to emulate his or her manifestation of aphasia. This was done by taking text [2], running it through ACES, and adjusting the software’s distortion parameters until the distorted text appeared ‘similar’ to the transcripts that were be generated by said individual. Only the sliders on the ACES interface were adjusted (no code was edited).

Once a model was set, every ‘original’ version of text sample generated by that individual was then run through ACES once. No text sample was repeated. This ensured that our study used whatever distortions ACES applied, without preference to more ‘successful’ distortions. These distorted sentences were then cleaned, fixing spacing or punctuation issues that may be a byproduct of removing or adding words. No word spellings, phrasing or other changes were made to the ACES text, further ensuring that the distortions shown to participants were precisely the ones generated by ACES.

[2] To remove bias, text used to calibrate distortions was unrelated to the aphasic transcripts used in this study.

	Text Sample Group	Participants' Label	
		Correctly	Incorrectly
Overall	Human	146 (50.69)	142 (49.31)
	Computer	155 (53.82)	133 (46.18)
	Total	301 (52.26)	275 (47.74)
Anomic	Human	87 (60.42)	57 (39.58)
	Computer	78 (54.17)	66 (43.83)
	Total	165 (57.29)	123 (42.71)
Agrammatic	Human	59 (40.93)	85 (59.03)
	Computer	77 (53.47)	67 (46.53)
	Total	136 (47.22)	152 (52.78)

Table 15.1: *Aphasia Turing Test Results*
Occurrence count with row percentages (accuracy) in parentheses

15.2.4 Population

We recruited 24 participants (3 male, 21 female) for inclusion in this study. Participants were students or faculty in Speech and Hearing Science Departments, as well as professionals in the Speech and Hearing Science community. We chose Speech and Hearing Science students, faculty and professionals as our target population because their training is specifically targeted towards the identification and treatment of speech disorders. Part of this training includes analyzing transcripts of conversations, diagnosing disorders based on language production (thereby distinguishing one from another), and treating the speech disorders themselves. We felt that this population was uniquely qualified to perform the discrimination tasks in this experiment.

We actively recruited from multiple institutions to cultivate a wide perspective on aphasia. Of our participants, all had taken at least one class that covered aphasia, and 67% of participants had personal experience with aphasia, or had taken a class that only covered aphasia. The population contained four current BS/BA students, 13 current MS/MA students, five participants with an MS/MA degree, and two participants with a PhD. The mean age of our participants was 26.4 (range 19 to 60 years).

15.2.5 Analytical Methods

To examine the quality of the ACES distortions, we compared the participants' responses to the 24 *Aphasia Turing Test* questions separately from the 24 responses to the *"How Human" Test* questions. For each Test, we treated all responses to that Test's questions as one uniform data set. Since a participant

contributes more than one data point within a test, the responses are correlated. Therefore, statistical tests must take into account the correlated nature of the data. For statistical comparison, we compared responses to text samples in the Human Group with responses to text samples in the Computer Group.This compares participants accuracy in distinguishing human distortions from distortions generated by ACES.

It is important to note that in this experiment, lack of statistical significance is the desired outcome. Statistically significant tests results generally, by definition, look for differences. If ACES distortions are indistinguishable from human generated distortions, we would see a lack of statistical significance ($p \geq 0.05$) between the Computer Group data set and the Human Group data set.

Responses to the *Aphasia Turing Test* were binary (users marked each text sample as Human or Computer in origin). This would suggest using a Pearson's Chi-Squared, Fisher Exact or Binomial test. However, these tests do not account for the correlated nature of the data (each participant answered multiple questions that were analyzed collectively). Generalized estimating equations (GEE) (Hardin and Hilbe, 2003) with a logistic regression[3] were used to account for these correlations. To augment our analysis, we also examined the percentage of data points that were labeled correctly, and the percentage labeled incorrectly. Lastly, we separated out the Anomic and Agrammatic text samples to determine if aphasia type impacted participants' ability to discriminate.

Responses to the *"How Human" Test* were categorical. This would suggest using a Two-Sample Wilcoxon Rank-Sum (Mann-Whitney) test, a more conservative metric than the Student's T-Test as it makes no assumptions about the data distribution. However, Rank-Sum tests do not account for the correlated nature of the data (each participant answered multiple questions that were analyzed collectively). Generalized estimating equations (GEE) (Hardin and Hilbe, 2003) with a linear regression[4] were used to account for correlation.

To further inform our analysis of the *"How Human" Test*, we also examined the distribution of data points. As an explicit measure of similarity of our two data sets, we utilized Rita and Ekholm's measure of similarity[5](Rita and Ekholm, 2007). This similarity metric utilizes a θ, or tolerance in the means between two data sets. We set a conservative θ to be one fifth of a Likert interval (0.2). This represents 5% of the possible answer range, and just over one eighth (13%) of the overall variance (1.50) in subject responses to the "How Human" Test Likert questions.

[3]Logistic regressions were used to test associations with binary outcomes (correct/incorrect labeling by participants).

[4]Linear regressions were used to test associations with scale responses (Likert scale 1-5) as outcomes.

[5]Rita and Ekholm measure uses a similarity limit, θ such that the difference in the averages of the two data sets is smaller than θ in absolute value. This can be determined by examining the 95% confidence interval for the difference between the two data sets. If the θ is greater than the right 95% confidence interval *and* $-\theta$ is less than the left 95% confidence interval, 'there is a difference' with $p<0.05$.

15.3 Results

Our data set consisted of 1152 observations (data points), from 24 participants. Of these, 576 observations were from the *Aphasia Turing Test*, and 576 were from the *"How Human" Test*. The following sections detail the quantitative results from our analysis.

15.3.1 Results for Aphasia Turing Test

As shown in Table 15.1, overall participants correctly discriminated Human vs. Computer slightly better than chance (52.26%). Similarly, within the two text sample Groups, participants correctly categorized slightly over 50% of the text samples. GEE tests indicated no statistical difference between subjects' ability to discriminate text samples from the Human group with text samples from the Computer Group (z= -0.75, p=0.46).

Further analysis of Anomic text samples (Table 15.1) shows similar results to that of the overall dataset. A comparison of the accuracy of rating the Human Group versus the Computer Group indicated no statistical difference between the two groups (z=1.06, p=0.29). Analysis of the Agrammatic text samples (Table 15.1) produced different results. Specifically, participant performance dropped considerably in their ability to correctly label text samples from the Human Group: 60% with Anomic text samples, 40% with the Agrammatic text samples (z= -2.08, p=0.04).

We performed a post hoc analysis (GEE test), comparing participants' performance between Anomic distortions and Agrammatic distortions within each text sample group (e.g., Anomic Human Group vs. Agrammatic Human Group) to determine if the participants could differentiate Anomic or Agrammatic better. Results showed no statistical difference between Anomic and Agramatic text samples from the Computer Group (z=0.12, p=0.91). However a highly significant difference was seen between Anomic text samples from the Human Group and Agrammatic text samples from the Human Group (z=3.28, p=0.001). This may indicate that the ability of participants to differentiate Agrammatic text samples that were from the Human Group (41%) was significantly poorer than when examining Anomic text samples from the Human Group (60%).

15.3.2 Results for the "How Human" Test

Table 15.2 shows summary statistics and sparklines for the distribution of participant responses to the "How Human" test, ranging from 1 (Definitely Human) to 5 (Definitely Computer). Overall,

	Text Sample Group	Mean (St. Dev.)	95% Conf. Int.	Histogram
Overall	Human	2.94 (1.22)	[2.80, 3.08]	
	Computer	3.05 (1.22)	[2.91, 3.19]	
	Total	3.00 (1.22)	[2.90, 3.10]	
Anomic	Human	2.73 (1.16)	[2.54, 2.92]	
	Computer	3.26 (1.21)	[3.06, 3.46]	
	Total	3.00 (1.21)	[2.86, 3.14]	
Agrammatic	Human	3.15 (1.25)	[2.95, 3.36]	
	Computer	2.84 (1.20)	[2.64, 3.04]	
	Total	3.00 (1.24)	[2.85, 3.14]	

Table 15.2: *Summary Statistics and Histogram Sparkline for "How Human" Test*
Response Range from 1(Distortions Definitely Human in Origin) to 5 (Distortions Definitely Computer in Origin). Histogram shows frequency of each Likert scale rating with 1 on the left, and 5 on the right.

participants rated ACES generated distortions as 3.05, showing a slight favor towards being computer in origin. Likewise, participants rated text samples from the Human Group as 2.94 overall, showing a slight favor towards being human in origin. However these slight shifts in preference showed no statistical significance ($z=-1.17$, $p=0.24$). Using the Rita and Ekholm's similarity measure (Rita and Ekholm, 2007), the two data sets were found to be statistically similar ($p<0.05$).

When we stratify our data by Aphasia Type, our results diverge. For text samples that had Anomic distortions, we observed a statistically significant difference ($z=-4.10$, $p<0.001$). Participants rated text samples from the Human Group as being more human (2.73) than text samples from the Computer Group (3.26). While these differences are about 1/4 of a Likert point away from a neutral score of 3.0, this result indicates that Anomic distortions were slightly less believable. This is confirmed with Rita and Ekholms' similarity measure ($p\geq0.05$).

Results of the Agrammatic text samples, however, ran contrary to ground truth. While there was a statistically significant difference between the Human Group and Computer Group ($z=2.32$, $p=0.02$), the mean responses were opposite to the origin of the text. Participants rated text samples from the Human Group as being more computer (3.15) than text samples from the Computer Group which were rated more human (2.84). These responses were biased in the wrong direction. It is also true that these results were statistically not-similar using Rita and Ekholms' similarity measure ($p\geq0.05$).

15.4 Discussion

In general, our results indicate that ACES provides a realistic set of distortions of aphasia. Our participants overall had difficulty differentiating between the origins of our text samples, and generally rated distortions as being right between definitely computer in origin, and definitely human in origin. Unlike most experimental setups, lack of significance is a positive outcome, validating the realism of ACES distortions. The remainder of this section discusses the specific results from each Test.

15.4.1 Aphasia Turing Test

Participants were unable to discriminate between distortions generated by humans with aphasia and distortions generated by ACES. In this regard, ACES distortions passed our variation of the Turing Test. With overall accuracies for both the Human and Computer Group hovering around 50% (nearly equivalent to random chance), and no statistical significance found between the two groups, we can conclude that ACES distortions are indistinguishable from those generated by humans with aphasia.

While the accuracy for identifying Anomic text samples from the Human Group rose slightly, the ability to correctly label Anomic text samples from the Computer group remained constant, and we saw no statistically significant difference between the Control and Human Group.

However, the results from Agrammatic text samples demonstrate an inability of participants to correctly identify text samples that originate from humans (41% accuracy). This probability is worse than chance, and is a statistically significant drop-off as compared to the accuracy for Anomic text samples. Further, the ability to correctly identify Agrammatic text samples from the Computer Group remained constant when compared to Anomic text samples(not statistically significant). Therefore we attribute the only statistically significant difference in the Aphasia Turing Test to participants' inability to correctly identify text samples from the Human Group, rather than an increase in their ability to identify text from the Computer Group. Taking this into consideration, we continue to see that participants had approximately a 50/50 chance of correctly labeling text samples from the Computer Group, still indicating that participants were unable to distinguish text samples from the Human and Computer Groups.

We can therefore conclude that, across the board, participants are generally unable to distinguish human distortions from ACES distortions, thus passing our variation on the Turing Test. This adds support to the claim that ACES creates realistic distortions of aphasia.

15.4.2 "How Human" Test

The overall results from the "How Human" Test paralleled those of the Aphasia Turing Test. Participants' ratings between Human and Computer Group showed no statistical difference. However, differences emerge when data is stratified by Aphasia Type. In general, Anomic distortions in the Computer Group tend to be labeled as more computer-like, while actual distortions in the Human group are correctly marked as being more human. Analysis confirms that this is a statically significant difference.

However, analysis of the text samples with Agrammatic distortions showed that participants generally believed that the ACES distortions were more human (2.84 on Likert scale 1-5), and the real text samples were more likely to come from a computer (3.15 on Likert scale 1-5). This difference was statistically significant. We therefore speculate on the possible causes of this surprising finding. First, our participants may have had difficulty in identifying Agrammatic aphasia. Second, transcripts (ours, or in general) may not have fully captured the nuances of Agrammatic aphasia. Third, the models and distortions ACES used were based on the same literature that is used to teach speech and hearing science students. It is possible that the literature does not fully describe the nature of Agrammatic aphasia. Therefore ACES may more closely match our participants' expectations of Agrammatic aphasia as compared to actual transcripts.

It is worth noting that mean scores (across Aphasia Type and Text Sample Group) are relatively close to the center of the 5 point Likert scale (equally human and computer). Examination of the distributions (last column of Table 15.2), reveals a single or double hump bell curve around a value of 3 on the scale. Thus indicating that participants generally were unable to categorize a text samples' errors as 'definitely' human or computer in origin.

Upon further examination, we determined that no one user performed notably better or worse when answering the "How Human" Test, suggesting that the results were consistent across participants. We also examined participants' qualitative responses, at the end of the "How Human" Test, commenting on how they made their decisions. Surprisingly, participant responses were not consistent. One participant mentioned placement of pauses in sentences, whereas another participant relied upon how 'obvious' a semantic replacement was. However, no two participants mentioned the exact same aspect of speech as being a key informative factor. Moreover, many participants' responses contained a phrase similar to that of participant 23, *"I was really surprised by how realistic the distortions were to me."*

15.4.3 Future Work & Limitations

This work represents an important step forward in validating ACES, and it's impact. As there are many distinctive subtypes of aphasia, the Aphasia Turing Test should be repeated with each of them, to explore the ability of ACES to emulate each specific type of aphasia. This vein of research would also help guide future development of ACES distortions, and improve the quality of the requisite distortions.

In addition, results from the "How Human" Test highlight that participants find Anomic distortions generated by ACES to be slightly more computer than human. However the specific reasons are unclear given the variety of user responses to the general question "How did you make your decisions?" We therefore propose a future investigation into ACES distortions, focusing only on Anomic errors. This study would ask participants to justify their decision on each question, rather than prompt for one reflective statement at the end of the study. This may provide specific insight into why ACES distortions fail and/or succeed.

Given the surprising Agrammatic text sample results in the "How Human" Test, future investigations need to be conducted as to why ACES distortions appear more human, and real transcripts appear more computer-like. In addition, this test should be repeated to ensure that this result was not in error.

This work should be replicated with other populations (e.g. computational linguists, or professionals who deal with aphasiacs every day). Because this study did not express examine the correlation between "aphasia experience" (or knowledge of NLP) and performance, we cannot fully state the degree of prior knowledge necessary to uncover the computer generated text. Likewise, having longer text samples (or even interactive IM conversations) could also uncover other strengths or weaknesses of the ACES system.

15.5 Conclusion

Empathy and understanding from family members, friends, professionals and caregivers directly impacts the quality of life and quality of care of individuals with aphasia. To this end, Hailpern et. al. developed ACES, a system which allows users (e.g., caregivers, speech therapists and family) to "walk in the shoes" of an individual with aphasia by experiencing linguistic distortions firsthand. ACES' distortion model was directly based on the literature in the fields of Cognitive Psychology and Speech and Hearing Science. While results from our initial experiment illustrate that ACES increases empathy

and understanding of aphasia, the research in Chapter 14 did not explicitly validate the distortion model. This chapter has made several contributions to address this limitation.

First, this chapter shows that participants from the Speech and Hearing Science community, whose training is specifically targeted towards the identification and treatment of speech disorders, cannot consistently differentiate computer and human generated distortions. Second, from our investigation of the realism of ACES distortions, we discover that overall, both human and computer generated distortions appear equally "realistic." However, when stratified by type of aphasia, we can see that ACES' emulation of Anomic aphasia is slightly less realistic than ACES' emulation of Agrammatic aphasia. Third, by validating the distortions used in Hailpern's original experiment, this chapter strengthens the original chapter's findings, showing that the distortions experienced were believable approximations of aphasia. Lastly, by coupling the results of this chapter and those of the original study, we add support to the feasibility of other language emulation research targeting other language deficits.

Chapter 16

Write-it Do-it

It is unfortunately the case that many friends and family avoid interacting with individuals with aphasia because they do not understand the disorder, lack empathy, and simply find interaction to be difficult. This lack of empathy can "erode the social bonds that give life meaning," and greatly diminish quality of care in a professional setting (e.g. by doctors and nurses) (Liechty and Heinzekehr, 2007). In Chapter 14 we focused on building empathy and understanding for individuals with Aphasia as a means to positively improving conversational patterns/quality. However, building empathy for a conversational partner is only one way to impact and positively change a conversation.

In a broad sense, there are two ways to remediate conversation quality: improving/increasing empathy, and studying the language/linguistics of conversation so as to provide concrete changes to conversation patterns. Both approaches, in theory, lead to the same result. However, the first one is implicit (allowing the changes to come from modifying how your feel) while the second is explicit (focusing on actionable modifications to your speech patterns).

This chapter now explores the latter approach. Our goal is to demonstrate that ACES can be used to uncover language and communication patterns inherent in conversations with individuals with aphasia. Further, we aim to tie these linguistic and conversational changes to changes in conversation quality (objective and perceived). Therefore, the intent of this research is twofold. The first is providing Psychology researchers with a large corpus of conversations[1] (with quality of conversation from both perspectives logged in addition to meta-data on what was intended to be sent). Second, and perhaps more importantly, is extending the original ACES work by demonstrating that our system can be used to understand how distortions impact linguistics (focusing on syntax not semantics) and conversation quality. Through an exploration (at a high level) of the conversation logs themselves, we have uncovered numerous patterns in communication, and how some of them relate to the frustration, clarity, and other dimensions of conversation quality. Notably, our results illustrate that there are two distinct

[1] In contrast to having face-to-face conversations with individuals with aphasia, ACES is faster, requires less time and resources, and allows for all conversations to exhibit the "same" type of aphasic errors (allowing for smaller N studies).

and independent impacts of ACES distortions on conversation quality. Further, we have found that many of the linguistic changes that people employ mirror a known theory in the Psychology literature (adaptation theory) - adding weight to the power, utility and applications of the ACES system.

These findings were unearthed through an experiment with 96 participants (in pairs) resulting in two 30-minute IM conversations per pair. Each conversation was structured following existing practice in the Psychology Communication literature, Write it Do it study (Beun and Cremers, 1998; Brown-Schmidt and Tanenhaus, 2008) and analyzed across multiple high level dimensions of linguistics (largely syntactical) and conversation quality (perceived and objective). Because ACES yields transcripts of what was sent as well as *what was intended to be sent*, ACES generates a much larger corpus of data than is typically available to most researchers studying aphasic language. We therefor focused on the intended message sent, rather than the message *after* distortion. Further, we were able to examine perceived conversation quality because all participants using ACES had no inherent language impairments. We therefore had the ability to probe their perspectives and feelings on conversation quality, their role, and how productive the interaction was [2].

We begin with a review of the literature related to the examination of communication and linguistic changes. Using prior work to guide our study design, we next present our experimental and analytical methods. We then discuss our participants, followed by a high level presentation of the study results. Our detailed findings, and subsequent linguistic analysis are then presented, concluding with a discussion of the implications of this study. Finally, we discuss future exploration and testing of our findings.

16.1 Related Literature on Communication & Linguistics

Most people are unaware of these subtle but important changes to language during interpersonal communication. However, through a detailed examination of these language patterns over conversations, we can uncover and quantify these patterns. The examination of the patterns of interpersonal communication is not an invention of this research, nor the computer age. Many researchers in fields such as Communication, Linguistics, and Psychology have explored how and why people communicate. Further, research in Natural Language Processing (NLP) has been examining the ability of computer systems/algorithms to understand, interpret, and quantify natural language. While the work in *this* chapter is not Psychology, Linguistics, Natural Language Processing nor Communication research,

[2] When interacting with individuals that have aphasia, the subjects' challenges with language can make surveys, questionnaires, and oral debriefing a challenge.

it is incumbent upon us to understand other work that could leverage our findings and the broader topic of this Part of the Dissertation (ACES).

In Appendix B.1, we highlight many of the potential applications and fields that could leverage ACES logs, providing support for the need for a tool like ACES to provide data sets for researchers. Of particular note for this work, is Aphasia Adaptation Theory.

16.1.1 Aphasia Adaptation Theory

It is a well established theory that individuals with aphasia change or "adapt" their speech patterns to those of their conversation partners (Kolk and VanGrunsven (Kolk and Van Grunsven, 1985) which defined this, Hartsuiker (Hartsuiker and Kolk, 1998) with syntactic priming, and Kolk (Kolk, 1995) about adapting language to the words and phrases that are readily available). A great example that talks about aphasia adaptation theory, and provides many good examples is (Ruiter, 2008; Ruiter et al., 2010). This work goes into different types of corrections such as "preventative adaptation" and "corrective adaptation." While the aphasic distortions are imposed upon the subjects by their impairment, they can learn behaviors that change their speech patterns in constructive ways. Further, (Kolk and Heeschen, 1996) suggest that many of the "symptoms" or characteristic output of aphasia may not be the disorder, but the adaptation of the individual to the impairment. This project with ACES allows us to directly examine how both the conversation partner AND the aphasic change their language. If researchers had logs of individuals experiencing aphasia, and adapting their language, they could uncover and further support the existing literature. In theory, tools like ACES could be used to generate robust datasets that capture both intended and received messages. If ACES does cause individuals to augment or adapt their language, it provides support for another potential use of ACES type systems.

16.1.2 Language Analysis in Human Computer Interaction

The examination of communication, language and textual communication has permitted the field of human computer interaction, most notably in the field of **C**omputer-**M**ediated **C**ommunication (CMC). Issues of trust in computer/textual and other forms of communication have been widespread (Bos et al., 2002; Wilson et al., 2006; Toma, 2010). While much of this work has not examined the language, rather just the outcome of the interactions. Scissors (Scissors et al., 2009; Scissors et al., 2008) is a noteworthy exception that examined alignment of text and mimicry.

Broadening out from issues of trust in CMC, Nguyen et. al (Nguyen and Rosé, 2011) examined idea "reflection" in online social communities by examining the occurrence (and reoccurrence) of common vocabulary words. The focus in this work is on semantics (word meaning) rather than syntax (word type or sentence structure)[3]. Their worked used the Kullback-Leibler (KL) divergence metric to examine the vocabulary space as it changed over 60 weeks. Wang et. al. (Wang and Fussell, 2010) examined CMC over IM, specifically by hand-coding messages as ideation, strategy, responses, (dis-)agreement, explanation, picture or other. Their focus was on the examination of responsiveness so as to calculate conversational distance (also using the KL divergence metrics for distribution of categories). Leshed's PhD. dissertation (Leshed, 2009) examined features from LIWC and the SYMLOG framework (Bales et al., 1979) to explore dimensions of interpersonal behaviors on which people interact over IM.

16.2 Research Questions

As discussed above, the intent of this research is twofold. The first is providing Psychology researchers with a large corpus of conversations, and the second is to demonstrate that ACES can be used to understand how distortions impact linguistics (focusing on syntax not semantics) and conversation quality.

To guide the second goal of this research (and the remainder of this chapter, study design and analysis), we scope this project by formulating several research questions. These questions are based on the wealth of literature discussing the theory and practice of the linguistics of interpersonal communication, allowing us to focus on demonstrating the quality impact of ACES distortions, and showing that participants *do* change their language in the presence of distortions.

1. How does conversation quality change in the presence of aphasic distortions?
2. What linguistic patterns occur as individuals adapt to aphasic distortions?
3. Which linguistic characteristics are universal, and which are dependent on a specific distortion?
4. Which linguistic characteristics, if any, are correlated with conversation success?
5. How do these conversational patterns evolve over the course of two, 30 minute conversations?

To this end, we scope this investigation as an initial exploration of the communication patterns that occur when individuals use ACES. We aim to produce a data set for future research in Computational

[3]It should be noted that the authors did briefly examine LIWC and POS though not a lot of focus went into that analysis and implications

Linguistics and Psychology, and demonstrate that there exist qualitative and linguistic changes that should be explored by experts in Psychology, Communication, and Computational Linguistics.

16.3 Experimental Design

In order to observe the effects of using ACES and Aphasic distortions on language and communication, we conducted an in-depth user study. Ninety-six individuals (cohorted in subject-pairs) engaged in two, 30 minute IM conversations with each other. Each **session** consisted of the following steps:

1. Demographic Questions
2. Conversation with Partner
3. Questions about Conversation
4. Conversation with Partner
5. Questions about Conversation
6. Final Set of Questions About Entire Experiment

The structure of our first study (Chapter 14) was a debate. However, to ground the conversation prompts in the psychology literature and ensure a lively conversation, we conducted a "**w**rite **i**t - **d**o **i**t" design (**WIDI**), following the work of (Beun and Cremers, 1998) and (Brown-Schmidt and Tanenhaus, 2008). One participant will be assigned the **Writer Task**, the other **Doer Task**. The Writer would be given a completed structure built out of various construction material (ranging from toy Legos to popsicle sticks or pipe cleaners). The Doer will have all the required pieces needed (plus extras[4]) to build the structure. The conversational goal would be for the writer to explain how to assemble the structure to the doer within the time frame. At the end of the first conversation, users will switch their tasks. If the structure is not completed within the 30 minutes, participants are stopped and their progress is recorded and scored. Further we can use this score as a metric to determine the team with the "best" structure, and provide that team with a reward[5]. This monetary prize further motivates participants to be good communicators. Because participants had two conversations, we constructed two different structures (see Section 16.5).

For our analysis, we need to be determine if the linguistic changes that impact conversation quality are unique to the aphasic conversations, or general principals. Participant pairs were therefore equally

[4] Participants will be told there are extra pieces.
[5] Participants will be told, at a high level, the metrics used to compute the "best" structure

Conversation 1		**Conversation 2**
Participant 1/Participant 2 + Structure		*Participant 1/Participant 2 + Structure*
AW/TD + Structure #1	⇒	TD/AW + Structure #2
AW/TD + Structure #2	⇒	TD/AW + Structure #1
AD/TW + Structure #1	⇒	TW/AD + Structure #2
AD/TW + Structure #2	⇒	TW/AD + Structure #1
TD/TW + Structure #1	⇒	TW/TD + Structure #2
TD/TW + Structure #2	⇒	TW/TD + Structure #1

Table 16.1: *Fully Counterbalanced Permutations*
AW = Aphasia Writer, **TD** = Typical Doer, **AD** = Aphasic Doer, **TW** = Typical Writer

divided into one of three cohorts a **Control Cohort** (where subjects did not experience aphasic distortions), a **Aphasic Writer Cohort** (where one participant had the writer task and experienced aphasia distortions), and a **Aphasic Doer Cohort** where one participant had the doer task and experienced aphasia distortions). Unlike our original, more traditional, experiment (e.g., Chapter 14), the aphasia cohort was split into two cohorts. During this task, the onus for contributing information is larger on the Writer (as compared to the Doer). To take this into account, and control for any bias the experimental structure may introduce, we split the Treatment Cohort. As a result, we can ensure to have an equal number of control conversations, conversations where the writer has aphasic distortions, and conversations where the doer has aphasic distortions (or the users' **Role**).

To fully counterbalance the treatment cohorts in this study, and control for order effects, we will need multiples of 6 pairs of conversation partners (see Table 16.1). Therefore, the 96 participants (48 conversation pairs) were divided up into three cohorts of 32 participants per condition (16 conversations per condition). The size of each cohort will therefore mirror the size of each cohort in the initial study (though the structure, and conversation task are vastly different). Participants were remunerated $25 for their participation the experiment (lasting approximately 90 minutes). A $30 bonus was awarded to the winning team in each cohort.

All participants (control and both treatment cohorts) experienced the same experimental protocol, as they were given the same prompts (only differing on whether or not aphasia and its distortions was explained as part of the protocol). Subjects were blind with respect to what other conditions existed within the experimental context. Both subjects in any given subject-pair were from the same same cohort (Treatment/Control), and remained within the same cohort throughout the experiment. Further, participants were told their task, and the task of their partner and whether or not they or their partner would experience aphasic distortions. Logs of IM conversations were taken by ACES (this includes the messages typed, distortions (if any) applied, and the resulting messages sent by

participants).

16.4 Types of Aphasia

We leveraged the models created during the Aphasia Turing Test (Chapter 15) to ground this experiment. Because none of the profiles tested in the Aphasia Turing Test experiment were extreme, they can be considered "typical" cases of aphasia. We therefore believe that they are good profiles to use for this experiment because we believe that they are realistic (to the extent that they cannot be distinguished by the SHS community) and they are grounded in real people's text distortions. We therefore used a profile based on an Anomic subject named Wolf from the unpublished data files used in (Menn et al., 1998a; Menn et al., 1998b) given to us by which were provided to the researchers by Lise Menn, University of Colorado. The implemented model was set to the following levels:

- **Distortion of Words –** Overall Correctness Slider: 75%
- **Distortion of Inflection –** Inflection Morphology of Verbs: 2%
- **Distortion of Function Words –** Dropping Function Words: 2%
- **Distortion of Non Fluency –** Omissions: 7%
- **Distortion of Non Fluency –** Semantic Description: 6%
- **Distortion of Other –** Pauses: 14%

16.5 Structures Used

Figure 16.1 and Figure 16.2 present pictures of the two structures used during the WIDIT experiment. Each structure was composted of a foam brick, with other objects attached to it. To ensure equal complexity (and vocabulary), both structures utilized legos, popsicle sticks, pipe cleaners, paper, stamps, and a pinwheel. However, the number, color, quantity, and placement vary greatly between the two structures. For easy distinction, we will refer to the structure in Figure 16.1 as the **Horizontal Structure** and the structure in Figure 16.2 as the **Vertical Structure**.

When designing the structures, we wanted to ensure an appropriate level of difficulty, so that participants would take at least 30 minutes to complete the task (so as to have long enough chat logs to perform our analysis). To this end, we recruited two engaged couples (friends of the researchers) to attempt the WIDIT protocol with early versions of the structures. By choosing pairs of participants

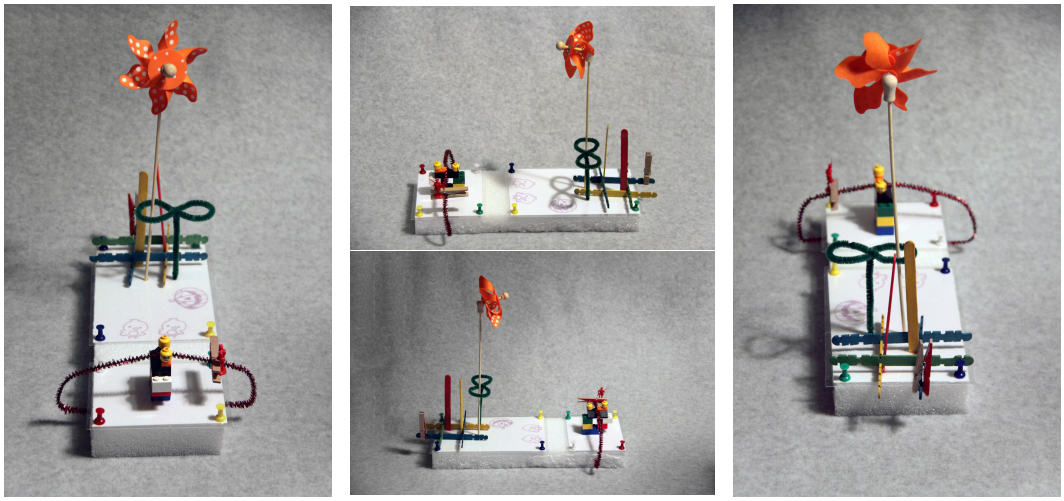


Figure 16.1: *Horizontal Structure*

that, in theory, had well established communication patterns, we could examine the "worse case" scenario for our task. Further, we could then discuss with the couples where the largest hangups were in the structures, and adjust the structures accordingly. The final structures used in the WIDI experiment are those shown in Figure 16.1 and Figure 16.2.

16.6 Measures of Conversation Quality

One goal of this study is to illustrate how having conversations using ACES impacts conversation quality. Given the structure of the WIDI experimental design in addition to questionnaires given to participants, we are able to quantitatively measure the IM conversational quality. Conversation quality can be measured in two important and different ways[6]. First is **Objective Quality**, or a measure of the effectiveness of the conversation that stays constant and unchanging (within the allowable error) across the persons measuring. These measures must not take opinion of subjective judgment into account. The second measure is **Perceived Quality**, or how each conversation partner felt about their experience or conversation. This technique dispenses with the objectively measurable features in favor of more intangible issues of assessing performance.

[6] In many ways, the division between Objective and Perceived quality mirror that of Subjective and Objective measures (Kirvesoja, 2001). We can therefore treat Objective Quality and an Objective measure, and Perceived quality and a Subjective measure.

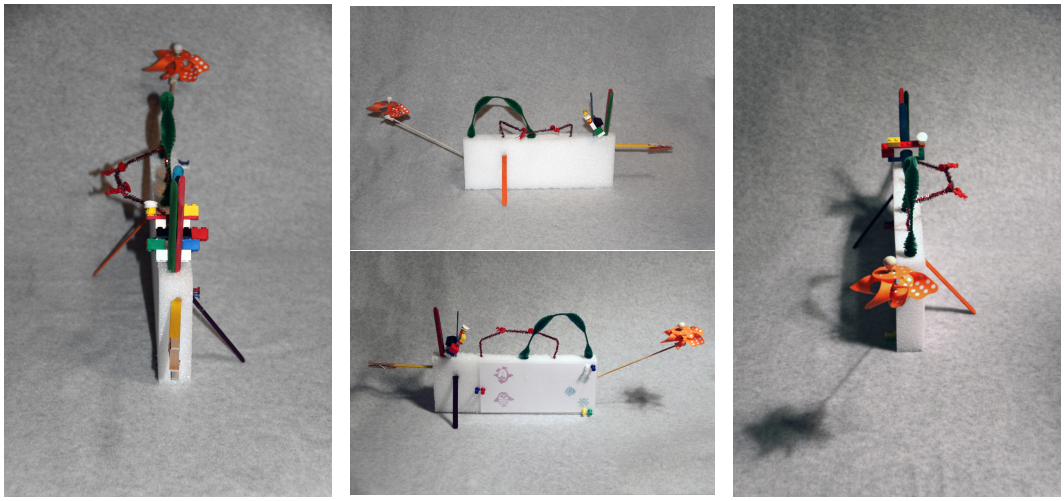


Figure 16.2: *Vertical Structure*

Measuring both aspects of conversation quality is important and central to this type of research (Kirvesoja, 2001; Cushman and Rosenberg, 1991). Not only can we uncover the relationship between objective and perceived quality (if any), but *how* users felt during their conversation is just as important as *what* was accomplished. Consider the real-world application of this research – improving conversation quality. If individuals with aphasia, and their conversation parters (family, friends, doctors, etc) *feel* less stress or anxiety, then conversations become easier and more common, improving quality of life. While having more objectively productive conversations, directly relates to how much people can accomplish and communicate.

The remainder of this section addresses the objective and perceived measures recorded. All measures conversation quality used are well accepted and published metrics within the scientific community.

16.6.1 Objective Measures of Conversation Quality

As mentioned above, we can measure the completeness of the built structure as a measure of quality. The "better" or more productive the conversation, will be closer to being complete than the less successful conversations. Each structure will be scored on completeness and accuracy. For both structures we created a list of all the connections it has. A **connection** is where one piece touches another, or how one piece is deformed/shaped/placed. A connection can be graded on both placement and orientation (where applicable). For every piece put on a structure which does not belong, will will deduct 1 point. Both structures are of the same relative complexity. The horizontal structure has 57

connections, while the vertical structure has 58 connections. To ensure cross structure comparisons we can use percentage completion rather than raw score.

We can also use time as a measure of conversation quality. If participants think they completed the structure within 30 minutes. If they did not. And if they think they did complete it, how far under 30 minutes were they.

16.6.2 Perceived Measures of Conversation Quality

One well accepted measures of conversation quality is the **I**owa **C**ommunication **R**ecord (**ICR**) by Duck (Duck et al., 1991). This is a very thorough set of questions across multiple dimensions (change, value, quality, control, conflict, and variability) of assessing and comparing conversations. In particular, we utilized the conversations that assess conversation quality. Participants are requested to describe the quality of their conversation on 10 dimensions, each using a 1 to 9 likert scale:

- **Relaxed** (1) to **Strained** (9)
- **Attentive** (1) to **Poor Listening** (9)
- **Formal** (1) to **Informal** (9)
- **Smooth** (1) to **Difficult** (9)
- **Guarded** (1) to **Open** (9)
- **Great Deal of Understanding** (1) to **Great Deal of Misunderstanding** (9)
- **Free of Communication Breakdowns** (1) to **Laden with Communication Breakdowns** (9)
- **Free of Conflict** (1) to **Laden of Conflict** (9)
- **Interesting** (1) to **Boring** (9)
- **You came away satisfied** (1) to **You came away not satisfied** (9)

A mean value from these 10 questions is calculated, and used as the ICR measure of conversation quality. **The lower the ICR score, the better the conversation.** Duck's original paper (Duck et al., 1991) presented expected outcomes for different segments of the populations (e.g., gender, relationship type and days of the week). We can therefore use these scores for comparison.

Another potential set of measures is the **I**nterpersonal **C**ommunication **S**atisfaction **I**nventory (**ICSI**) by Hecht (Hecht, 1978) which is intended to investigate a person's reactions to a conversation. Unlike the ICR, the ICSI presents the respondent with 19 statements that they must agree or disagree with (on a 7 point likert scale). However, because the measure is intended to assess spontaneous conversations,

some questions did not relate to the WIDI experimental design. We therefore asked the following subset of 10 questions from the ICSI:

- Q1:The other person let me know that I was communicating effectively
- Q2:Nothing was accomplished.
- Q3:I would like to have another conversation like this one.
- Q5:I was very dissatisfied with the conversation.
- Q8:The other person showed me that he/she understood what I said.
- Q9:I was very satisfied with the conversation.
- Q11:I did NOT enjoy the conversation.
- Q12:The other person did NOT provide detail for what he/she was saying?
- Q14:We each got to say what we wanted.
- Q16:The conversation flowed smoothly.
- Q18:The other person frequently said things which added little to the conversation.

In the original paper by Hecht, the author uncovered three factors that can be extracted from the questions:

- **Factor 1:** Affect/morale - *Questions 2, 3, 5, 9, 11 and 16*
- **Factor 2:** Substance/salience - *Questions 18 and 19*
- **Factor 3:** Free Interaction - *Questions 13 and 15*

Given the subset of 10 questions asked, and the relevance of each of the three factors to the WIDI design, we utilized Factor 1 for our ICSI analysis. Much like the ICR, responses are scored (following the scoring protocol in the original paper) and then a mean is calculated and used as the ICSI measure of conversation quality.Because some statements are "positive" (something good happened) and some are "negative" (something bad happened), the scoring of negative questions (e.g. Question 2) are flipped, so that positive interactions will always have a higher value. **The higher the ICSI score, the better the conversation.** *This is in contrast to the ICR which uses lower scores as the better outcome.*

16.6.3 5 Measures of Conversation Quality

For each conversation, we will collect a total of 5 measures of conversation quality:

1. Objective Conversation Quality
2. Writer's ICSI

3. Writer's ICR
4. Doer's ICSI
5. Doer's ICR

There is one assessment of Objective Conversation Quality. This Objective Score is based on the completed structure, and is the result of the conversation between *both* partners (each partner effectively shares the Objective score). For each of the measures of *perceived* conversation quality (ICSI and ICR) there are two assessments made (one by each partner). The participant in the Writer Task records an ICSI and ICR score, and the participant in the Doer Task also records an ICSI and ICR score.

16.7 Linguistic Features and Measures of Language

One way to examine the conversation logs is to extract linguistic features of conversation. Using basic NLP and IR techniques, we can extract both simplistic (e.g. lines/message sent) and more complex (e.g. Parts of Speech, Pragmatics, Questions) linguistic features (largely syntactic) of conversations. These features can provide insight into how people communicate (Control Cohort) and how participants adapt their conversation style to adjust for the distortions cased by ACES. To this end, we extracted 6 parts of speech (Function Words, Adjectives, Verbs, Nouns, Adverbs, and Content Words), utterances or lines, use of punctuation, as well as occurrences of subject questions, and pragmatics. From these linguistic features (POS, Questions and Pragmatics), we can examine their raw occurrences (e.g. number/count of nouns) and their percentage occurrence (e.g. % of all words that are nouns).

It is important to know that these linguistic features are not intended to be a comprehensive list of all linguistic analysis that could be performed. As this is research in Computer Science, not Computational Linguistics, our goal is to illustrate the feasibility and usefulness of the ACES system by demonstrating that there are linguistic changes that occur due to Aphasic Distortions. A complete and comprehensive examination of all linguistic (as well as semantic and sentiment) impacts is reserved for researchers in other communities with a more complete and throw understanding of these issues.

16.7.1 Parts of Speech

Parts of Speech (POS) are an incredibly important aspect of understanding language, especially for natural language parsing and information retrieval (Jurafsky et al., 2000). Showing changes in POS is necessary to demonstrate potential use of more advanced NLP and IR techniques for uncovering

more complex linguistic structures. In addition, POS and number of utterances/messages are accepted measures for analyzing conversations and have been used in the past to examine aphasic speech (Saffran et al., 1989).

16.7.2 Questions

Questions are some of the main types of dialogue acts that occur in task-oriented dialogue (like the WIDI study) (Jurafsky et al., 2000; Allen and Core, 1997; Carletta et al., 1997; Core et al., 1998) focusing on requesting information, or confirming information is correct. While we are not delving into many of the subtypes of questions, the presence of questions can indicate requests for repair (implying problems in understanding) (Jurafsky et al., 2000). We focused on identifying common questions (see Appendix B.2.1) that can be modeled with context-free rules(Jurafsky et al., 2000) plus any message that ends with a question mark .

16.7.3 Pragmatics (Continuers)

Pragmatics[7] help us understand "how discourses are structured, and how the listener manages to interpret a conversational partner in a conversation." (Jurafsky et al., 2000) While the term "pragmatics" can have a broad and varying definition, in this project, we specifically examine "continuers," also referred to as "backchannels" or "acknowledgment tokens," which, in the context of a task-centric dialogue, can be thought of as "discourse questions." "Continuers" form a subset of the five main types of Pragmatics (Clark and Schaefer, 1989). Such utterances could be a user stating *Sure* or *Mmm Humm* (see Appendix B.2.2), and are used to help signal the conversation partner to continue, stop, or when they have achieved the target goal (Jefferson, 1984; Schegloff, 1982; Yngve, 1970) (central to the WIDI experiment design).

16.8 Other Measures

In addition to the above measures of conversation quality, we also asked participants associated questions about their conversations. These questions focused on strategies used, what they found

[7] How to find and measure pragmatics is a complex and open area of research. We use a simplistic approach of finding and counting the occurrence of known pragmatic phrases for this study, while acknowledging that a wider variety of approaches can and should be used in the future.

effective, what they "wish" their partner would have done. This qualitative responses allow us to "get inside the head" of participants, and understand what they did, and why. This data, in conjunction with our robust data set of log files, can further aid researchers analyze our logs with the potential to draw connections between what people said they did, and what actually occurred when examining the conversational logs.

16.9 Participants

All 96 participants did not know each other, and were assigned randomly to a Cohort and Task. At the beginning of each session, we asked participants a series of demographic questions. Table 16.2 contains the mean and standard deviation for participants in each cohort. In addition, we tested to see if our cohorts' makeup war statistically different (see Section 16.10 for more details). Results from this analysis are included in Table 16.2. There were no statistical differences in age, gender, educational attainment, or prior knowledge of aphasia.

16.10 Statistical Methods

The scalar nature of our quantitative measures would suggest using a Two-Sample Wilcoxon Rank-Sum (Mann-Whitney) test, a more conservative metric than the Student's T-Test as it makes no assumptions about the data distribution (normal or otherwise). However, Rank-Sum tests do not account for the correlated nature of the data: each pair of participant had two conversations, and their interactions will clearly be correlated. Generalized estimating equations (GEE) (Hardin and Hilbe, 2003) with a linear regression[8] were used to account for correlation.

To test for correlations, we likewise used Generalized estimating equations again to account for the correlation in data. In correlation analysis, we can examine the coefficient to have an indication of the slope of the regression (and the direction of correlation). The coefficient is not the same as an R^2 (correlation coefficient), and should not be treated as such. Therefore, the GEE tests' coefficients are not as easy to interpret as a Pearson's Coefficient. We therefore *also* ran a pairwise correlation coefficient using the Pearson's Correlation test. Sadly, Pearson's Correlation test does not account for the correlated nature of the data, so we do not report the p-value, and the calculated R^2 value should

[8] Linear regressions were used to test associations with scalar responses as outcomes.

	Control Cohort (C)	Aphasic Writer Cohort (AW)	Aphasic Doer Cohort (AD)	p-value C vs AD	p-value C vs AD	p-value AW vs AD
Age	30.75 (10.76)	29.75 (9.41)	27.00 (11.69)	0.70	0.15	0.29
Sex (% male)	46.88	43.75	50.00	0.80	0.80	0.61
Educational Attainment †	2.00 (1.30)	1.94 (1.22)	1.44(1.37)	0.84	0.08	0.12
Knows what Aphasia is∗	43.75	46.88	31.25	0.80	0.80	0.61
Taken Class on Aphasia∗	21.88	25.00	12.50	0.80	0.80	0.61
Worked with Aphasia∗	3.12	6.25	0.00	0.80	0.80	0.61
Family/Friend with Aphasia∗	3.12	3.12	0.00	0.80	0.80	0.61

Table 16.2: *Write-it Do-it Participants*
Continuous variables presented as mean (SD), p-values calculated by GEE
∗ Represents % that respond affirmatively
† Education attainment scores were 0=High School, 1=AA, 2=BS/BA, 3=MS/MA, 4=PhD

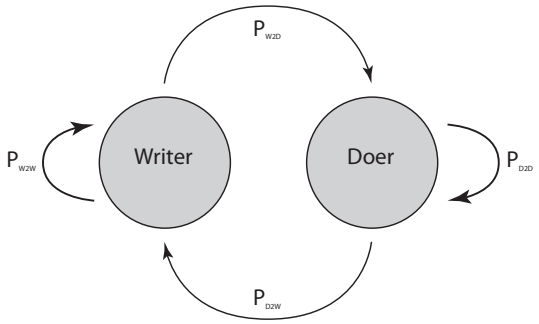


Figure 16.3: *Write it Do it Turn Taking Hidden Markov Model / Finite State Machine*

only be treated as an indication of the correlation magnitude (and not a hard-and-fast 'true' value). R^2 values range from -1.0 (negatively correlated) to 1.0 (positively correlated).

By having a fully counterbalanced study design, we minimize the impact of learning/order effects. However, for a robust analysis, we also examine the impact of order/learning through a GEE analysis, comparing both conversation quality, and linguistic measures.

As a final statistical analysis, we examine the turn taking between subjects (writer and doer). To perform this analysis, we can treat a conversation as a Hidden Markov model (HMM) or Finite State Machine (FSM)[9] (see Figure 16.3. By examining the occurrence of turn taking, we can explore:

- Turn Taking or Command/Answer – Writer to Doer (P_{W2D})
- Turn Taking or Acknowledgement/Question – Doer to Writer (P_{D2W})
- More (Amount/Complex) Problems – Doer to Doer (P_{D2D})
- Expanding/Clarifying/Continuing – Writer to Writer (P_{W2W})

The probabilities are calculated by taking the count of the transitions (e.g. Writer to Doer) divided by the count of all pairs that *start* with the same sender (e.g. Writer). Thus the probabilities are calculated as follows where N is the mathematical symbol indicating count, W is messages from a writer, and D is messages from a doer:

- $P_{W2D} = \frac{N_{W2D}}{N_W}$
- $P_{W2W} = \frac{N_{W2W}}{N_W}$
- $P_{D2W} = \frac{N_{D2W}}{N_D}$
- $P_{D2D} = \frac{N_{D2D}}{N_D}$

[9] In this instance, both HMM and FSM terms are accurate description of the model being constructed

It should be noted that P_{W2D} + P_{W2W} and P_{D2W} + P_{D2D} sum to a probability of 1.0. Subsequently, by knowing the outcome of one of the probabilities within a pair, it is intuitive that the other is 1-P. Therefor when reporting, we will only report P_{W2W} and P_{D2D}.

16.11 Results

Our analysis can be broken up into two main sections: Conversation Quality and Linguistics. The former examines the impact of ACES distortions on conversation quality (and the interrelationship between measures of conversation quality). The latter attempts to illustrate that there is a linguistic effect of ACES distortions, thus highlighting the potential applications of ACES-like solutions for other researchers. In addition, we examine order/learning effects. For cohorts with distortions, all analysis examines the *intended* messages (message before distortion by ACES), rather than the message *after* distortion. This allows us to examine how users adjusted their own language to accommodate or adapt to the Aphasic errors.

16.11.1 Conversation Quality

The mean, standard deviation, median and inter quartile range of the 5 measures of conversation quality (see Section 16.6.3) are presented in Appendix Tables B.1, B.2 and B.3. Each table presents summary statistics for one of the three cohorts. Overall, the Control Cohort had the "best" conversations, followed closely by the Aphasic Doer Cohort, while the "worst" conversations were had by the Aphasic Writer Cohort. When we compare the conversation quality between the three cohorts (see Appendix Table B.4), we see that across all 5 measures, there is a statistical difference between the conversation quality in the Control Cohort and the Aphasic Writer Cohort. We also see several strong differences between the Aphasic Writer and Aphasic Doer Cohorts (though not in the ICSI measures). Interestingly, there was no statistical difference been the conversation quality of the Control Cohort and the Aphasic Doer Cohort. In addition, there was no statistical difference in perceived quality between the writers and doers (see Appendix Table B.5) within each of the three cohorts.

In theory, the 5 measures of conversation quality should all reflect the same thing: the quality of the conversation. To examine the inter-relationship between these five measures, we completed a correlation matrix (see Appendix Tables B.6, B.7, and B.8). Each table presents summary statistics for one of the three cohorts. Surprisingly, the only two consistent correlations observed were between the

two measures of *perceived* quality. There is a negative correlation between Writer's ICSI and Writer's ICR, and there is a negative correlation between the Doer's ICSI and Doer's ICR. Having a negative correlation is expected, because while the higher the ICSI outcome is the better the conversation, the ICR uses lower scores to signify better conversations. Further, these correlations had a high effect size (see Appendix Tables B.9 - B.11). It's interesting to note that there was *no* significant correlation found between Objective conversation quality and any of the Subjective conversation quality measures.

16.11.2 Linguistic Measures

16.11.2.1 Raw Linguistic Measures

The summary statistics for the raw linguistic measures are presented in Appendix Tables B.13, B.14, B.15, B.16, B.17, and B.18. Each table presents summary statistics for one of the three cohorts, with the first three focusing on the Writer's linguistics and the second three focusing on the Doer's linguistics. Comparative statistics are presented in Appendix Tables B.19 and B.20. Calculated correlations between measures of conversation quality and linguistics are presented in Appendix Tables B.22-B.39.

There is a clear statistical difference in the number of words (both Writer's and Doer's) across all three cohorts. As people simply talk more (number of words increase), the occurrence of each part of speech will likely increase as well. We tested the correlation between total number of words and each Part of Speech in turn (including line count and punctuation count), and found a high degree of positive correlation (with $p<0.01$) [10]. Further, we saw a correlation between questions asked (Writer's and Doer's) and word count ($p<0.01$) in all three cohorts[11]. These findings call into question the efficacy of using any raw count of part of speech values and the basis of analysis. We may still, however, analyze the number of lines to examine how often a participant spoke, and the number of words to determine how much was said (in quantity, not quality).

16.11.2.2 Percentage Linguistic Measures

An alternative to raw occurrence of linguistic characteristics (e.g. number of nouns) is to use percentage occurrence (e.g. percentage of words that are nouns). This metric is robust to the amount of dialogue

[10] For all pairings except Control Cohort – Writer's Punctuation ($p=0.75$) and Aphasic Writer Cohort – Writer's Punctuation ($p=0.04$).

[11] With the exception of Writer's Questions in the Aphasic Doer Cohort ($p=0.75$). We did *not* see any correlation between Writer's or Doer's continuers in the Control or Aphasic Doer cohorts. However, there was a highly correlated relationship between Writer's and Doer's Continuers in the Aphasic Writer Cohort ($p<0.01$).

	Control to Aphasic Writer			Control to Aphasic Doer			Aphasic Doer to Aphasic Writer		
Writer % Function Words	0.43	0.40	▼	0.43	0.43	–	0.43	0.40	▼
Writer % Content Words	0.57	0.60	▲	0.57	0.57	–	0.57	0.60	▲
Writer % Adjectives	0.12	0.13	–	0.12	0.11	–	0.11	0.13	▲
Writer % Verbs	0.13	0.11	▼	0.13	0.13	–	0.13	0.11	▼
Writer % Nouns	0.28	0.32	▲	0.28	0.27	–	0.27	0.32	▲
Writer % Adverbs	0.05	0.04	▼	0.05	0.05	–	0.05	0.04	▼
Writer % Questions	0.06	0.05	–	0.06	0.09	–	0.09	0.05	▼
Writer % Continuers	0.06	0.04	▼	0.06	0.07	–	0.07	0.04	▼
Doer % Function Words	0.39	0.40	–	0.39	0.38	–	0.38	0.40	–
Doer % Content Words	0.61	0.60	–	0.61	0.62	–	0.62	0.60	–
Doer % Adjectives	0.10	0.11	–	0.10	0.10	–	0.10	0.11	–
Doer % Verbs	0.18	0.16	▼	0.18	0.20	–	0.20	0.16	–
Doer % Nouns	0.28	0.29	–	0.28	0.27	–	0.27	0.29	–
Doer % Adverbs	0.05	0.04	▼	0.05	0.04	–	0.04	0.04	–
Doer % Questions	0.31	0.49	▲	0.31	0.28	–	0.28	0.49	▲
Doer % Continuers	0.29	0.24	–	0.29	0.28	–	0.28	0.24	–

Table 16.3: *Comparative Statistics Between the Three Cohorts – Mean Values and Change Direction in Linguistic Measures as Percentages*
Values and arrows represent statistically significant change/direction, dashes indicate non-significant change.
Arrow indicates direction of change from the first to the second number
Blue is significance $p \leq 0.05$, Red $p \leq 0.01$

produced, and the subsequent analysis will focus on the differences between the language usage. The summary statistics for the raw linguistic measures are presented in Appendix Tables B.41, B.42, B.43 for the Writer's Linguistics and Appendix Tables B.44, B.45, B.46 for the Doer's Linguistics. Each of the three tables presents summary statistics for one of the three cohorts.

We compare the Writer and Doer's linguistics between the three cohorts in Appendix Tables B.47 and B.48. This data shows a clear difference between the Aphasic Writer Cohort and the Control Cohort (with similar differences when comparing the Aphasic Doer Cohort to the Aphasic Writer Cohort). In the Writer's language, we see an increase in Content words in the Aphasic Writer cohort (as compared to the Control or Aphasic Doer Cohorts), most notably with an increase in nouns (though verbs, adverbs and continuers all decrease). In the Doer's language, we see a slight drop in verbs and adverbs in the Aphasic Writer Cohort (as compared to the Control Cohort). We also see a very large increase in the percentage of questions asked between the Control/Aphasic Doer Cohort and the Aphasic Writer Cohort. We further compare the Writer's Linguistics to the Doer's Linguistics in Appendix Table B.49. We see statistically significant differences between the Writers and Doers within each cohort, most likely illustrating the differences in their Roles.

	Objective	Writer ICSI	Writer's ICR	Doer's ICSI	Doer's ICR
Writer % Function Words	–	D	–	–	–
Writer % Content Words	–	D	–	–	–
Writer % Adjectives	–	–	–	C	–
Writer % Verbs	–	–	–	–	–
Writer % Nouns	D	–	–	–	–
Writer % Adverbs	CD	–	–	C	CW
Writer % Questions	–	W	–	–	–
Writer % Continuers	–	D	–	D	–
Doer % Function Words	–	C	W	–	–
Doer % Content Words	–	C	W	–	–
Doer % Adjectives	–	CD	C	C	C
Doer % Verbs	CW	–	–	–	C
Doer % Nouns	CW	–	C	–	–
Doer % Adverbs	–	D	–	–	–
Doer % Questions	–	–	–	–	–
Doer % Continuers	C	C	C	C	C

Table 16.4: *Significant Correlation Was Found Between Linguistic Changes and Conversation Quality*
C = Control Cohort, W = Aphasic Writer Cohort, D = Aphasic Doer Cohort
Blue is a positive correlation, Red is a negative correlation

We then examined the correlations between conversation quality (Objective and Perceived) and Linguistic Changes. These are presented in Appendix Tables B.50 through B.79, though summarized in Table 16.4. There are clearly some statistically significant correlations, that have a modest effect sizes.

16.11.2.3 Learning Effects

Over the course of a conversation (and between conversations) users have the potential to learn new "strategies" for improving their conversation quality. By fully counterbalancing all comparisons in the above analysis (equal number of first and second conversations in each group during a statistical comparison), it is interesting to examine the impact of learning/order on conversation quality and use of linguistics. The summary statistics are presented in Appendix Tables B.81 - B.85 for the first conversation and Appendix Tables B.87- B.91 for the second conversation. The statistical GEE comparison is presented in Appendix Table B.92. For simplification, Appendix Table B.93 summarizes the Appendix Tables by illustrating when there was a statistically significant change, and the direction of said change.

The most noteworthy change (or absence of change) was in the five measures of conversation quality. No measure, in none of the cohorts, changed between the first and second conversation. In the Control

Cohort, the percentage of Writer's questions decreased while the Doer's verbs increased. Within the Aphasic Writer Cohort, the Writer's language changed, with function words decreasing (thus increasing content words). This change mostly came from increasing the use of adjectives and nouns. It is also worth noting that the writer asked fewer questions in the second conversation. Within the Aphasic Doer Cohort, the Doer generally decreased the number of words used, cutting the number of questions asked and increasing pragmatic statements. The Writer also decreased the number of continuers used.

16.11.2.4 Turn Taking

We present the summary statistics and statistical analysis of the turn taking probabilities in Appendix Tables B.94 - B.100. Across all cohorts, the turn taking behavior remained consistent with the Writer sending multiple messages just under 50% of the time, and the Doer generally sending one message at a time. It should be noted that the variance was quite large in all three cohorts. Further, turn taking changes had no correlation with any change in conversation quality.

16.12 Discussion

At the highest level, we can see that our experimental design was a success. The manipulation of conversations with ACES has a clear and statistically significant impact on both objective and perceived conversation quality. By observing that objective conversation quality is impacted, we have shown that Aphasic distortions in conversation do impact the output/goal of the conversation. Further, distortion of the language in a conversation also impacts how people perceive the conversation (as observed by ICSI and ICR scores)[12].

While a significant difference in conversation quality is observed between the Control Cohort and the Aphasic Writer Cohort, we do not see a difference between the Control Cohort and the Aphasic Doer Cohort (see Appendix Table B.4). This observation opens up a new question; is the Shannon Information (Shannon and Weaver, 1962) from the Doer to the Writer not as important (so therefore any distortion does not matter), or is the "amount" of information that needs to get across from the Doer to the Write so much smaller/easier to understand (e.g. short pragmatics like "ok", or "got it"), that the distortions have little impact on the information conveyed? When we compare the number of

[12] While we do not see statistical significance for ICSI scores between the Aphasic Writer and Aphasic Doer Cohort, this may be due to statistical power.

lines and words used by Writers and Doers (see Appendix Table B.40 for comparisons and Appendix Tables B.13-B.18 for raw counts), we see a striking difference with the Writer statistically producing more lines and many more words (around 3 times as many). It is outside of the scope of this analysis to explicitly measure Shannon Information in each message, so we therefore cannot assess if the Writer has more (in quantity), or more complex, information to convey. However we see this as a promising future research question.

16.12.1 Measures of Conversation Quality

We had originally theorized that all five measures of conversation quality represent the same "thing," implying that users would perceive the conversation as having been better or worse in correlation with the success of the task. However, Appendix Tables B.6, B.7, and B.8 show a different story: there is no relationship between Objective Conversation Quality and Perceived Conversation Quality. Yet, we do see a statistically significant negative impact of Aphasic Distortions on all forms of Conversation Quality[13] (see Appendix Table B.4).

When these two analysis are considered in concert, we can conclude that distortions of aphasia (by ACES) impact conversation in two distinct, yet independent, ways (illustrated by Figure 16.4), by affecting the the product (or goal) of the conversation, as well as the interpersonal interaction. These conclusions are clearly grounded on the quality of the measures used. While the Objective Conversation Quality is, by definition, objectively observable, perceived quality is not. However, by having two measures, that have a high effect size correlation ($R^2 > 0.6$) and is highly significant ($p < 0.05$) with each other (Appendix Tables B.9, B.10, and B.11 for effect size and Appendix Tables B.6, B.7, and B.8 for significance), we can infer that they are valid and verifiable measures. This supports the conclusion that Aphasia (or at the very least ACES' distortions which simulate Aphasia) impacts both how we feel about the interaction as well as the conversation output quality/goal.

Another conclusion based on the analysis of the multiple measures of Conversation Quality is that there are two aspects of conversation that can be targeted for "improvement." We can attempt to improve the information conveyed to generate a more productive output (Objective). We can also attempt to change how people feel and react to challenging interactions (Perceived). While both are important, improving perceived conversation **must** be targeted, because only improving the objective output itself may not improve the perceived quality (since the interaction will still be challenging). Given

[13] While we do not see statistical significance for ICSI scores between the Aphasic Writer and Aphasic Doer Cohort, this may be due to statistical power.

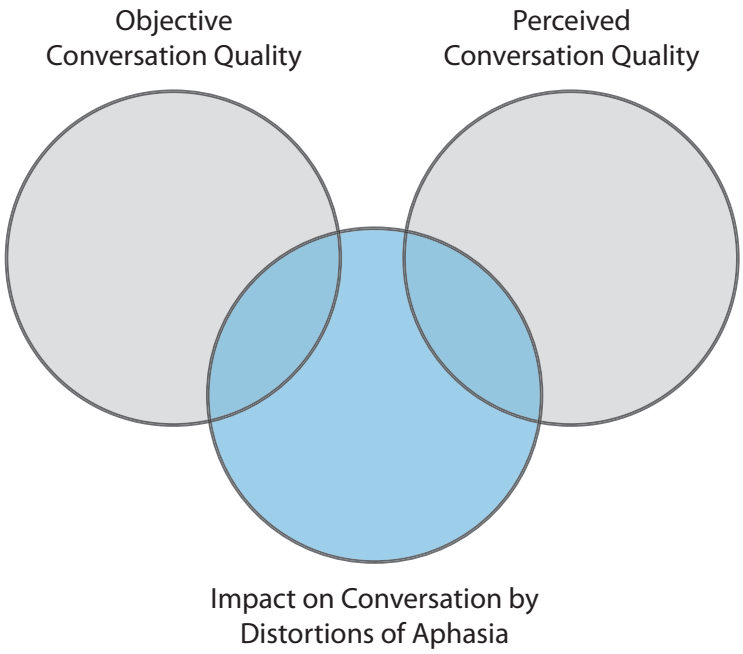


Figure 16.4: *Venn Diagram of Conversation Quality and Distortion Impact*

the previously discussed work on Improving empathy, understanding and patience for conversation partners (see Chapter 14) focuses on augmenting and changing how people *feel*, we hypothesize that by providing users (and potential conversation partners, clinicians, and/or doctors) with the experience of speaking with aphasia (through ACES) we can effectively improve the perceived conversation quality. This would need to be verified in a future experiment.

16.12.2 Linguistic Effects

We observed a strong impact of Aphasic distortions on several linguistic measures. This clearly indicates that users were actively altering their language in an attempt to overcome the distortions of ACES. As shown in Appendix Tables B.19 and B.20, both Writers *and* Doers decrease their number of words when *their* text is distorted. However, they do not change the number of lines sent. This indicates that when a subject has their text distorted, they are more concise with their communication, though not less infrequent.

We therefore must examine what changes within those messages result in a decrease in words. When Writers had their text distorted, they increased the percentage of words that were nouns (see Table 16.3)

while sacrificing verbs, adverbs, continuers and function words. This suggest that Writers adopted a more curt method of communication, focusing on the object rather than the action. Further, when the Writer's text was distorted, Doers greatly increased the number of lines that contained a question (see Table 16.3) signaling an increased need for clarifications or questions [14]. This observation potentially has broader implications. Participants are behaving the way that real non-fluent aphasics behave. This strengthens the arguments for the theory of aphasic adaptation – patients know that it's hard to talk, and so they concentrate on producing shorter sentences (Kolk and Heeschen, 1990) with high-information words, such as nouns (Salis and Edwards, 2010), and dropping low information words, like function words, verbs, adverbs, etc (Thompson et al., 1997; Kolk and Van Grunsven, 1985).

It is interesting to note that there were no differences between the Control Cohort and the Aphasic Doer Cohort in linguistic changes (Writer and Doer). This further suggests that the impact of the distortions in the Aphasic Doer Cohort was minimal at best.

16.12.3 Linguistic Effects on Conversation Quality

Further, we examined (Table 16.4) the correlations between linguistic changes and conversation quality. It does appear that when the Doer uses more nouns (in the Control and Aphasic Writer Cohorts) that Objective Quality increases whereas the Doer using more verbs (in the same Cohorts) decreases Objective Quality. This suggests that clarifying the object of a discussion is important whereas the action is not.

When the Doer is distorted, the Writer has a higher conversation quality (ICSI) when the Writer uses more function words (providing structure, more eloquent and less curt sentences). However, when the Writer is distorted, the Writer has higher conversation quality (ICR) when the Doer uses more function words (providing structure, more eloquent and less curt sentences). This suggests that the Writer (who does the majority of the communication) is happier when the individual without the distortions (whether himself or his partner) sounds less curt, and more natural. In other words, even if someone has distorted text, their partner should try to sound as "normal" as possible rather than mimic or curtail their own speech[15].

In the Control Cohort, both the Writer and the Doer dislike the Doer using adjectives, and likes the Doer using continuers (across all perceived conversation quality). Further, when the Doer uses more

[14] An area of future research would be a deep examination of the types of questions Doers asked, and how (if at all) the Writer responded).

[15] This is consistent in the Control Cohort where the Writer has higher conversation quality (ICSI) when the Doer uses more function words

continuers, the conversation quality increases. These observations suggests that the Doer's use of continuers (acknowledging task completion and message comprehension) should be studied further.

Overall the scattered differences are hard to reconcile into an overarching pattern of behavior. This may suggest the somewhat surprising conclusion that people have few consistent, intuitive approaches to improving communication in conversation, and that merely letting people try whichever approaches strike them will not result in any consistent improvement. This is supported by the observation that many of the patterns we uncovered (correlations with conversation quality) appear within the Control Cohort, and not within the Cohorts with distortion [16]. In other words, people are not necessarily very good conversationalists by nature, when placed in a stressful conversational situation. This implies that researchers should attempt to provide family members, friends, doctors, speech pathologists, and other care givers strategies to aid in communication. ACES can, in this experimental context, provide a testing platform to uncover the *true* impact of these strategies on language (do people do what they were tasked), objective and perceived quality.

16.12.4 Order/Learning Effects

Discussing order/learning effects is convoluted because there are so many pronouns as uses switch Tasks and Roles between conversations. To simplify, we situate our discussion within a fictional session with subjects Alice talking with Bob and Charles talking with Dave. In the first conversation, Alice is the Aphasic Writer and Bob is the Typical Doer. In the second conversation, Bob is the Aphasic Writer and Alice is the Typical Doer. For Charles and Dave, Charles is the Typical Writer and Dave is the Aphasic Doer in the first conversation. In their second conversation, Charles is the Aphasic Doer, and Dave is the Typical Writer.

The first observation (see Appendix Table B.93) we see in terms of "learning" is that the most learning happened by the Task with the distortions (when Bob takes over Aphasic Writing from Alice and Charles takes over Aphasic Doing from Dave). The user that "changed" their language was on the other side of the distortions first (Bob and Charles), and adjusted their language when they switched Tasks. This is in contrast to Alice and Dave who used the same language as their partners (Bob and Charles) when they took over the Typical Roles. Therefore, this data asks an interesting question: was

[16] An alternative explanation for not seeing these patterns in the Cohorts with distortions is due to, perhaps, lack of statistical power (not enough users altered their continuers, adjectives or adverbs). To test these theories, a future investigation should be done, where users are explicitly tasked (when distortions are present) to increase and decrease adjectives, adverbs and continuers. This would allow us to study the impact of these correlations in the presence of errors.

this a change by the subject based on what they "wish" their partner had done (or not done) in the first conversation, or an augmentation having been frustrated during the first conversation?

We observe that within the Aphasic Writer Cohort, the Doer that becomes the Writer (Bob) focuses their communication on providing details, nouns (the objects) and descriptions, while refraining from asking questions. We see a similar change in questions for the Writer that becomes the Doer in the Aphasic Doer Cohort (Charles). In both cohorts, there seems to be a desire to refrain from asking any questions. This may be a reflection of a "get it done" approach to the task. Users may be trying to exchange as much information as possible with their partner, rather than clarifying ever small detail. However, given the lack of change in conversation quality, these changes do not appear to have any impact on how well the conversation fared.

16.12.5 Turn Taking

When we consider the Turn Taking between users (see Appendix Tables B.94 - B.100), we observe a consistent pattern with Writers dominating conversations regardless of cohort. This is interesting in that, people did not attempt to become more or less frequent in communication when their text, or their partners text, was distorted. This lines up very well with the lack of change in number of lines produced across cohorts (as discussed above). Further, there were no correlations between changes in Turn Taking (which varied greatly) and Subjective or Objective conversation quality.

16.13 Future Work

In many regards, this analysis is the tip of a much larger examination that should take place. While our analysis highlights that ACES has a distinct impact on language and conversation quality, our examination is preliminary at best. To this end, we envision 3 distinct avenues of future work to expand on this project and delve deeper into understanding and mini zing the impact of Aphasic distortions on conversation quality and linguistics.

First, a followup study should be done (following the same protocol) that tests a strategy for improving conversation quality. Using one cohort's condition (e.g. Writer with Aphasic Distortions), give half of the participants a known strategy to employ (e.g. ask more questions) and allow the other cohort to do what comes naturally. This will show whether that strategy for improving conversation quality works, and how it impacts (in practice) user's language.

Second, given that this data is two fold (both intended and distorted messages), we believe that there is a rich analysis that is yet to be conducted, examining the relationship between the language of the non-distorted subject, and the *distorted* messages sent. How do these users align, and adjust? How does the non-distorted user clarify the distortions?

Third, we strongly believe that there are many more levels of linguistic analysis that should be performed. As this is not linguistic or psychological research, these additional analysis are well outside the scope of this project. As we highlight earlier in this chapter, an examination of communication accommodation (e.g. LIWC), alignment[17] (structure and content), and aphasia adaptation. These analysis are far more complex than the above examinations, and many are coded (by hand) by experts in language. Similarly, a deeper examination of the question/responses that occur (how often users ask for clarification, and how often questions are answered) is a worth examination that can only be done by hand.

Fourth, a question that arrises is the cause of the above correlations. Rather, are the correlations we observed a result of the conversation quality being good/bad, or the driving factor in changing the conversation quality. As the study was designed, we are not able to answer these questions since we only sought to uncover correlations that naturally occur in conversation. To assess cause and result, further investigations are needed that experimentally test precise causation hypothesis.

Lastly, it would be intriguing to explore the performance of individuals with Aphasia doing the WIDI task. This would be an important direction to add further support to the generalizability of our findings.

16.14 Conclusions

This analysis has uncovered that ACES is a remarkable system that can be used to examine the impact of language distortions on conversation quality and linguists. ACES' distortions of Aphasia have a clear and negative impact on conversation quality (Research Question 1). Through a detailed examination, we can clearly see that these effects are quite complex, and reach far beyond the superficial intuition that distortions are "bad." We have shown that ACES distortions impact different aspects of conversation quality, and, more surprisingly, that these impacts (on objective and subjective quality)

[17] It should be noted that there may be some intrinsic challenges in performing an alignment analysis on this data set. Specifically, unlike a debate style conversation or DayTrader game as used by Scissors (Scissors et al., 2009), participants in the WIDIT study had extremely different tasks to complete. As we have seen this greatly impacts the type and amount of language used. This, therefore, would greatly impact structure and content. Further, as users switch from object to object on the structures, their language content (vocabulary) will shift quite quickly. This may cause further challenges to align vocabulary that is used only a handful of times.

are not correlated. This surprising finding stresses the need for research into improving how users perceive their interactions (building empathy) and providing strategies that can directly impact the outcome of conversations. Perhaps the most exciting finding is that participants appear to be behaving (in their linguistics) the way that real non-fluent aphasics behave (acceding to Adaptation Theory). This, if studied in more detail, has many ramifications for ACES, and this approach to understanding language disorders.

Further, this research has shown that ACES is a powerful tool for understanding how distortions impact language. Even a high level examination of parts of speech, continuers and discourse have shown that the introduction of Aphasic distortions impacts language (Research Question 2). And yet, these changes are not universal (Research Question 3), nor do they correlate with conversation quality (Research Question 4). This is a very important finding, suggesting that people do not instinctively know how to improve their conversations in distorted interactions. This implies that **both** individuals with aphasia **and** their conversation partners need external guidance to improve their interactions. Having reported these findings at an expository level, a deeper investigation by experts in computational linguistics, communications, and psychology is needed to understand and address more complex interpersonal dynamics in the presence of Aphasic distortions. However, it is now clear that ACES provides a clear platform (and a rich dataset) for researchers across many disciplines.

Chapter 17

Individuals with Aphasia Study Design and Feasibility

One of the key applications of ACES is to provide students/clinicians/family a tool with which to increase empathy and awareness to the challenges in communication experienced by individuals with aphasia. ACES is the first tool created, that we are aware of, which allows a neurologically individual to experience the communication distortions of aphasia . This has many applications. For example, while using ACES speech pathology students can practice communication skills with an individual with aphasia and also learn patience, empathy and understanding. Experience such as this is far more cost-effective and easier to facilitate than utilizing multiple individuals who have aphasia (which require remuneration and dedication of time) for hands-on experience. In this area alone, ACES provides a major contribution to improving empathy and understanding among speech pathology students.

An important area for future work with respect to the impact of the Aphasia Characteristic Emulation software is to test ACES in a real-world context. Questions we might consider are:

- Does exposure to ACES impact performance on school examinations for speech pathology students studying aphasia?
- Does ACES improve communication between speech language pathologists and individuals with aphasia?
- Similarly, does ACES improve communication between family members/friends and those with aphasia?
- What is the duration and frequency of exposure to ACES required to improve communication?
- And for all of the above questions, how does an intervention with ACES compare to more traditional educational tools (i.e., reading information on aphasia or watching informational videos about individuals with aphasia).

All of these questions would be important to address in future work. Thus, the focus of this chapter is on outling research involving speech language pathologists and individuals with aphasia in a clini-

cal/educational setting. Specifically, it provides two key contributions to solving the above questions. First, we outline a potential and plausible study design. Second, we highlight and discuss one of the challenges that accompany designing any study of this type of real-world application– statistical power. By outlining future work for ACES, we hope to lay the ground work for future investigations that will take into consideration the limitations we detail.

17.1 Study Motivation

To validate the real world application of ACES as an education tool, we outline a series of proposed studies that could be conducted with the aim of exploring how exposure to ACES might impact speech pathologists and family/friends. Specifically we propose a study design the explores communication between speech pathologist students and individuals with aphasia. The potential applicability of ACES as an educational tool (in a classroom setting) was echoed by participants in our Empathy Study (Chapter 14).

This chapter therefore provides a study design which examines the applicability of ACES as a supplement and/or compliment to traditional educational techniques (e.g. reading information on aphasia or watching videos of individuals with aphasia) we would hope to see that ACES provides an important insight that allows communication parters to be more empathic and better conversation partners.

17.2 Study Measures

To date, ACES research on empathy has focused on self-perceived change (i.e., how much does the user feel that their attitudes and perceptions have changed). However, one of the primary applications of ACES is to improve interpersonal communication between individuals with aphasia and students/clinicians/family by increasing and promoting empathy. Because the accepted measures of "empathy" are more holistic, they cannot be used to do a before/after within subject study design. Our research on communication and linguistic changes (Chapter 16) provides measures of conversation quality (many of which do relate to empathy and understanding). We therefore look to these same metrics for validating the impact of ACES as an educational tool to improve conversation quality.

17.3 Study Design

Participants would all be masters students in speech and hearing science, working towards an advanced clinical degree in speech pathology. Ideally, all participants would have the same exposure to aphasia in their prior education (all having taken, or not taken, a course whose sole focus was aphasia). Because the masters degree in Speech Pathology allows those students to practice speech therapy, this educational intervention is directly targeted at this population.

Participants would be categorized into two groups (i.e., A and B). The groups would represent the two educational interventions. **Group A** would be educated with ACES, **Group B** would be educated with printed material (e.g. text books, transcripts, professional papers, or perhaps even video) typically found in an undergraduate or graduate Speech and Hearing Science Class on aphasia. This allows us to examine a more "active" intervention (ACES) with more "passive" interventions (traditional classroom study material). Clearly, an alternative "active" intervention might be exposure to individuals with aphasia. However in a real-world context, such an intervention would be both financially expensive and time consuming. It would also require a large number of individuals with aphasia to dedicate their own time to work with each student one-on-one. As a result, a one-on-one interaction with individuals with aphasia is not a practical real-world intervention and is therefore omitted from this study proposal.

In addition to the student participants, we would also recruit a small number of individuals with aphasia (around one quarter to one fifth of the subject pool size). Persons with aphasia would serve as our dependent measure.

Interventions would be administered over multiple weeks as follows (with participants in each group receiving their respective intervention).

Step 1: Students complete an initial questionnaire which include basic demographic questions and Mehrabian's measure of empathy test (Mehrabian and Epstein, 1972). This serves as a comparison to ensure/test that the three groups of students (and individuals with aphasia) have comparable empathy levels upon study entry.

Step 2: All student participants have a face-to-face conversation with each of the individuals with aphasia. Conversations last 20-30 minutes and discuss a non-controversial topic. Following the conversation, the student and the individual with aphasia complete post-conversation questionnaires utilizing The Iowa Communication Record (Duck et al., 1991) and questions

from The Interpersonal Communication Satisfaction Inventory (Hecht, 1978). This serves as a baseline measure of the student's interaction with individuals with aphasia.

Step 3: Each week, students participate in the study three times (Monday, Wednesday and Friday) for 40 minutes per session. During the 40 minute sessions, student participants receive the specific intervention for their group (e.g., receiving print material and reading it, watching video, or having ACES conversations with another student participant). At the end of each session, student participants answer reflection questions on what they "learned" during the session. By "learned" we ask them to self reflect upon their experience, and determine what value they received.

Step 4: Every third week (2 weeks post-intervention) students have a checkpoint, another face-to-face conversation with an individual with aphasia. The two measures are re-administered.

Step 5: The above steps are repeated for 12 weeks (the duration of a school semester). In total, student participants will have 8 weeks of intervention and 4 checkpoints.

At the conclusion of the study (12 weeks), analysis of data can be performed. We would use a between subject ANOVA to compare the impact of the different interventions on the communication and conversation quality between speech pathology students and individuals with aphasia.

17.4 Feasibility and Statistical Power

When designing a study, one of the primary concerns to a researcher is **Statistical Power**. Statistical power is the probability of rejecting a false null hypothesis (what is the probability of detecting significant differences should they be present). Power should be close to one, and in general, studies should have at least 0.8 (80%) power. Statistical power has a direct relationship between sample size (number of participants), number of groups (or conditions) and effect size (the difference between the means of the different groups) [1]. One way to determine statistical power is with the use of power and sample size software. we used NCSS PASS (statistical software) and the above scenario to set some given parameters, and solve for others. Given the above study design, the number of groups is set at 2, and following standard HCI study design our α (the probability of rejecting a true null hypothesis) is 0.05.

[1] Effect size dictates standard deviation

Power	N per group	Total N	Std. Dev.	Effect Size
0.07	10	20	5.00	0.10
0.14	10	20	2.50	0.20
0.25	10	20	1.65	0.30
0.40	10	20	1.25	0.40
0.56	10	20	1.00	0.50
0.72	10	20	0.83	0.60
0.85	10	20	0.71	0.70
0.92	10	20	0.625	0.80

(a) *Solve for Power - Two Cohorts*

Power	N per group	Total N	Std. Dev.	Effect Size
0.80	394	788	5.00	0.10
0.80	100	200	2.50	0.20
0.81	44	88	1.65	0.30
0.81	26	52	1.25	0.40
0.81	17	34	1.00	0.50
0.81	12	24	0.83	0.60
0.80	9	18	0.71	0.70
0.84	8	16	0.625	0.80

(b) *Solve For Sample Size - Two Cohorts*

Table 17.1: *Parameter Power Analysis using NCSS PASS - Two Cohorts*
Values solved for are in Red, manipulated effect size values are in Blue, and values held constant are in **Black**
Due to the relationship between effect size and standard deviation, we highlight both values in Blue.
NCSS PASS set mean(Group 1) = 1, mean(Group 2) = 2,mean(Group 3) = 3

Generally HCI experiments have approximately 20-30 participants. Given $\kappa = 2$ (number of groups), we might assume 10 participants per group (20 total). If we expect to see a 10% effect size, our study would have a statistical power of 0.07%. If we continue to adjust the effect size (Table 17.1a) we continue to see a very underpowered study. It is only an effect size between 60% and 70% would result in the study having the minimal acceptable power. While ACES is a novel and effective tool, we don't foresee ACES having this large of an impact on conversation.

An alternative is to explore the effect of changing sample size (N) while keeping statistical power constant. In other words, what is the number of participants needed to see a certain effect size at a given power. Once again, study design dictates that $\kappa = 2$ and by convention we set a power of 80% (0.80, the minimum acceptable power in published papers [2]). As Table 17.1b indicates, the study requires a very large sample size of 788 participants for an effect size of 10%.

17.4.1 Increasing Experimental Complexity

Our above experimental design utilizes two groups for comparison. However, we can envision adding a third group as a control group or separating the video from reading material as a third intervention type. While this is a reasonable study design choice, it raises addional concerns over limited statistical power. Assuming three cohorts ($\kappa = 3$), we re-run the calculations presented in Table 17.1. The updated findings of Power and N are presented in Table 17.2.

Assuming we expect to see a 10% effect size, our study would have a statistical power of 0.07%. Continuing to adjust the effect size (Table 17.2a) reveals a very underpowered study. It is only when an effect size of 60% is reached that our study has the minimal acceptable power. While ACES is a novel and effective tool, we don't foresee ACES having this large of an impact on conversation.

When detecting a smaller effect size, sample size (N) could be increased in order to keep statistical power constant. In other words, what is the number of participants needed to see a certain effect size at a given power. Once again, study design dictates that $\kappa = 2$ and by convention we set a power of 80% (0.80, the minimum acceptable power in published papers [3]). As Table 17.2b indicates, to see an effect size of anything less than 40%, the study requires a very large sample size (930 participants for an effect size of 10%).

[2] While the value of power was set to 0.80 in NCSS PASS, the software does ensure equal participants per group. Therefore, the actual calculated value of power can vary a bit to ensure multiple of 2 participants in the sample size. Power values must therefore be at least 80%

[3] While the value of power was set to 0.80 in NCSS PASS, the software does ensure equal participants per group. Therefore, the actual calculated value of power can vary a bit to ensure multiple of 3 participants in the sample size. Power values must therefore be at least 80%

Power	N per group	Total N	Std. Dev.	Effect Size
0.07	10	30	8.00	0.10
0.14	10	30	4.00	0.20
0.27	10	30	2.70	0.30
0.46	10	30	2.00	0.40
0.57	10	30	1.75	0.50
0.81	10	30	1.35	0.60
0.92	10	30	1.16	0.70
0.97	10	30	1.02	0.80

(a) *Solve for Power - Three Cohorts*

Power	N per group	Total N	Std. Dev.	Effect Size
0.80	310	930	8.00	0.10
0.80	79	237	4.00	0.20
0.81	37	111	2.70	0.30
0.81	21	63	2.00	0.40
0.81	16	48	1.75	0.50
0.80	10	30	1.35	0.60
0.83	8	24	1.16	0.70
0.86	7	21	1.02	0.80

(b) *Solve For Sample Size - Three Cohorts*

Table 17.2: *Parameter Power Analysis using NCSS PASS - Three Cohorts*
Values solved for are in Red, manipulated effect size values are in Blue, and values held constant are in **Black**
Due to the relationship between effect size and standard deviation, we highlight both values in Blue.
NCSS PASS set mean(Group 1) = 1, mean(Group 2) = 2,mean(Group 3) = 3

17.5 Feasibility and Scope within Dissertation

While performing this analysis, and validating the real-world impact of ACES as an educational intervention is a critical next step; our power analysis clearly demonstrates that the scale of this study would require extensive time, participants and resources to conduct. Given the scope of this dissertation document, and completed research presented on ACES herein, we hold that this study is outside the realm of feasibility for this document.

That said, this proposed study is important for future validation of ACES and integration of ACES into a standard Speech and Hearing Science curriculum. As researchers and educators begin considering ACES as an educational solution, we hope that they will not only document its quantitative impact, but also gather student's qualitative responses. Our goal is that this chapter will serve as an outline of how to proceeds.

Chapter 18

Summary of Findings

This part of the dissertation presents the motivation, background and theory related to the creation of ACES, a tool and model that emulates the effects of aphasia first hand. Based on the existing literature, we built ACES, a novel system that enables users to experience the speech-distorting effects of Aphasia firsthand. ACES deploys a probabilistic model (grounded in literature from Cognitive Psychology and SHS) that can distort a userÕs Instant Messages (leveraging techniques from NLP), transforming the original message into text that appears as though it had originated from an individual with Aphasia. Results from a quantitative evaluation, with 64 participants (consisting of experts and the general public), show that ACES strongly increased understanding, knowledge of and empathy for Aphasia. To further validate the ACES software and its emulation of Aphasia, we conducted a follow-up Turing Test study with 24 experts in Aphasia. My results show that experts, whose profession includes the diagnosis and categorization of this disorder, were unable to distinguish between text samples generated by ACES and those that came from individuals with Aphasia. Based on these successes, we conducted an extensive analysis of the impact of ACES distortions on conversation quality and linguistics. This study generated a large data set for future researches, and also highlighted the number of changes that aphasic distortions have on communication.

The implications of this thesis part are broad and far reaching. Aphasia, an acquired language disorder that affects over one million Americans, severely hindering their ability to communicate. While these individuals may appear to possess poor cognitive function, the problem resides in their language, not in their ability to think. Unfortunately, quality of life and care can suffer as friends, family, speech pathologists and doctors avoid interacting with Aphasic individuals, finding those interactions to be difficult. While roughly one million Americans may have Aphasia, the disorder impacts the lives of millions more. From a real world perspective, this research suggests that ACESÕs applications can, through clinical (friends and family) or educational (therapists and practitioners) use, improve quality of life and care.

Further, ACES is the first system to emulate the effects of a language disorder. It opens up a new avenue of HCI and CS research. From the perspective of a linguist or a psychologist, ACES can be used to study how humans adjust their conversational and linguistic structure due to the distortions caused by Aphasia. For technologists (CS and HCI), the tool is a contribution in its own right, having passed a Turing Test. ACES also illustrates a new approach to computationally studying language which is distorted due to language disorder, or unfamiliarity with that language (e.g., by emulating non-native speakers).

18.1 Limitations

During the initial ACES experiment, we examined empathy by asking participants to self-reflect on how the experience impacted them. However, we do not know the duration of the impact (how long did it last) or quantitatively measure the true empathy impact. While not asked, these are important questions that are left unanswered and should be explored.

While the Aphasic Turing Test conducted showed amazing results, this examination should be expanded to other user groups (e.g. computational linguists or psychologists). As individuals in Speech and Language Pathology are not experts in the ability of computers to distort language, their ability to judge the validity of the distortions may be somewhat limited. However, these other user demographic (while more in-tune to computer capabilities) may be unable to identify aphasia correctly. Similarly, the turning test experiment does not test users ability to differentiate ACES distortions (or real aphasia) from random textual distortions.

The final Write it Do it experiment, while on a larger scale, did not quantitatively highlight any changes an individual could make that would directly improve conversation quality. While we had hoped to uncover just such a natural change, our sample size may still be too small to uncover a successful change that only a small portion of the users chose to implement. Likewise, since the linguistic measures examined were generally at a lower level (and were not examined by hand), more complex changes (or ones that cannot be programmatically detected) were not analyzed.

18.2 Future Work

Based on our work with ACES, we believe that there is great potential for improving classroom education and cross-cultural business interactions by emulating the effects of **E**nglish as a **S**econd

Language (**ESL**). Many students and professionals are challenged and frustrated when teachers, faculty and collaborators do not speak fluent English. We hypothesize that many of these challenges may derive from lack of empathy and understanding for the challenges of communicating in a foreign tongue. Given the rich collection of ESL research and data from years of its implementation in grade school and higher education, we believe an ACES-like system could be created to improve cross-cultural business interactions and classroom education. The broad applicability of this research allows for multiple ESL emulations (e.g. mandarin to english, spanish to english) and multiple applications (e.g. businesses, grade-school, higher education).

Building on the models and approach created for ACES, we believe that a system that could "un-distort" Aphasic errors could be built. This system potentially could be used in real-time (e.g. IM or face-to-face) or in asynchronous communication (e.g. email). While we would likely never achieve 100% accuracy, we strongly believe that even moderate un-distortion (e.g., via general summary) could greatly improve day-to-day interaction for individuals with Aphasia. A similar system could be built that provides word selection aids or context summary tools to conversation partners, helping both individuals to choose their vocabulary, and understand the conversation.

A much broader direction of future work could be combining these language analysis and disjoin techniques with data mining and pattern analysis. In theory, if a system could find some patterns in specific individual's speech (who have language impairments) the system itself may be able to diagnose their disorder. Ideally this future system could perform diagnosis off of just a few things a person says (leveraging active learning) and create an accurate "impairment profile".

Much like ACES was designed to aid users by building empathy, and helping researchers by examining conversation quality and linguistic changes – we believe a conversation aid could be created to help users through conversations. Leveraging the "un-distortion" work suggested above, this conversation aid could provide suggestions to users of how to change their language to improve their conversations. This tool would effectively act as the grease in the wheels of interpersonal communication. When conversation falters, this system would be able to suggest remedies, or help implement corrections to errors made (by providing alternatives, or even "un-distorting" on the spot).

Part III

Conclusions

Chapter 19

Conclusions

This document has explore two distinct research agendas focusing on the intersection of langue and human computer interaction. The first part of this dissertation focused on the individual with an impairment; specifically, how HCI and CS technology can help develop language and speech production in children with autism and speech-delays. The second part focused on those around an individual with an impairment; in particular, discussing how HCI and CS solutions can increase empathy and understanding for individuals with aphasia. In addition, the second part of the examined how that same technology can uncover the impact of language impairments on linguistics and conversation quality for individuals with impairments and their conversation partners. These two parts of this dissertation are linked together by a common theme, understanding the relationship between the design of computer interfaces and individuals with an impairment, specifically individuals with communication impairments. The results of this work illustrate new directions of HCI research both in a purely academic community, and with implications for real-world applications.

The findings of the first half of this dissertation have direct applications to impact the lives of individuals with autism and speech delays. We developed SIP or the Spoken Impact Project that consists of three major thrusts of research; 1) A framework to facilitate video annotation, and an instantiation of said system called VCode and VData; 2) A^3, a set of coding guidelines that allow researchers to assess non-verbal subject interaction with computer based feedback; and 3) A detailed quantitative analysis of SIPS through the use of VCode, VData and A^3. Results from this analysis suggest that computer based feedback systems hold the potential to greatly impact the vocalization of non-verbal subjects. Further, the subsequent project, VocSyl, demonstrated the construction of a fully-functional tool to teach multi-syllabic speech skills using TCUID. The resulting software was well received by participants and their parents – with a iPad application to be released online for free. We hope that through this work we illustrate how HCI researchers can build, design and study tools to teach language skills to children with severe speech-delays, make them engaging, and impactful.

The findings of the second half of this dissertation have direct applications to researchers studying the effects of Aphasia and clinicians and family members working with those impacted directly by aphasia. To this end, we developed ACES, a tool and model that emulates the effects of aphasia first hand. These emulations were quantitatively shown to be realistic (by passing a turing test) as well as improve empathy and understanding of aphasia. We conducted an extensive analysis of the impact of ACES distortions on conversation quality and linguistics. This study generated a large data set for future researches, and also highlighted the number of changes that aphasic distortions have on communication. The implications of this half are broad and far reaching. Individuals who have friends, family or clients with aphasia can utilize ACES to help them empathize with their conversation partners. Likewise, researchers can use ACES to examine the linguistic and conversation quality impact of aphasia – potentially developing new strategies for improving communication.

At a higher level, this work illustrates that human computer interaction solutions can help improve communication for individuals with impairments. Specifically, this work examined how HCI and CS techniques can impact language development (critical for communication), empathy (critical for communication with individuals who have aphasia) and the role communication impairment has on linguistic patterns as well as conversation quality (to understand how impairments impact communication). Each of these broad themes (language development, linguistic patterns, and empathy/understanding) are equally important aspects of communication. And this work shows that within each of these themes, technology has a powerful role to play. Further, this work demonstrates that by exploring and researching technology that impacts those with impairments, we can advance our knowledge of the human condition, develop new computer solutions, and create technology that can improve the lives of all individuals – with a large and broadly reaching impact.

Appendix A

A^3 CODING GUIDE

The following is the coding guide distributed to coders. Included here is the full description of the behavior that must be observed to mark each variable. The variables are broken down into four passes, and the mode of video playback is also noted.

A³ CODER GUIDE

*denotes ranged event

PASS 1: *Use Standard Playback Mode*

*Non-Child Audio: Durations when audio/sound made & see on bars (you can hear it) and is NOT coming from the child /overlaps with child with child's sounds OR sound caught on volume bars, but is not identified as from child (unknown source). Include REGULAR heavy breathing. Mark whole segment if "contaminated"

PASS 2: *Use Interval Payback Mode (Continuous Playback)*

Smiling: When the child appears to be smiling during the past 3 seconds, and in the past 3 seconds, we at some point could see his face.

No Face: Can not see face (to determine smile) at all in the past 3 sec.

Oriented @ Screen: When the child, in the past 3 seconds, was facing towards the screen within 90 degrees. Spinning through the 90 degree arc should not be counted. Rather, time where the facing direction is within the arc for at least a "moment.

Auditory Focus: During the time frame, did the child get closer in proximity to, or touch the speaker. OR Child is not in the visual arc, in response to computer sounds, orient to the visual arc.

PASS 3: *Use Standard Playback Mode*

Use a 2 second pause between end and start to delineate between vocalizations
Also mark sounds even if not recorded by computer

*Laugh: The sound should NOT be able to be transcribed as a speech like vocalization. To qualify as laughter it needs to be paired with a positive affect.

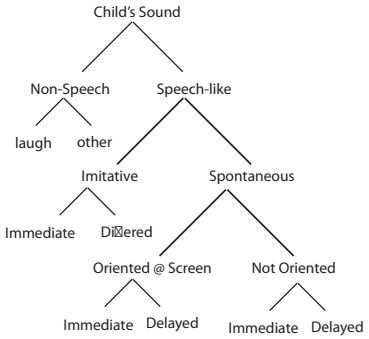
Non-Speech Vocalization: When a vocalization made is non speech, and is not a laugh. Includes gulps, screech, grunt, lip pops, ticks, heavy sighs, etc. Lasts for 2 seconds before code again.

*Speech Like Vocalizations: When a vocalization is made by the child, the phonetic construction of the sound should be noted as an event. Marked at the start of the sound
Use the annotation hotkey to write the sound made
New Sounds are formed by gap of 2 seconds, OR separated by a non-speech sound, or laugh, or computer sound (while the child is not making a speech-like sound).

Turn taking: the computer makes a sound, and then the child starts sound if nucleus of final syllable (vowel) has been initiated
Speech Like Sounds Only

BIGmack Switch: When a child presses the switch, mark this at the start of the press. Do not mark when switch is pressed by anyone else.

For each speech like vocalization mark one of the following Correspondences.



Spontaneous: creating a non-imitative vocalization

I+S spontaneous (Immediate, Oriented @ Screen, Spontaneous): while oriented towards screen (within in 2 seconds of starting sound), speaker, within 5 seconds of end of source sound

D+S spontaneous (Delayed, Oriented @ Screen, Spontaneous): while oriented towards screen (within in 2 seconds of starting sound), speaker, after 5 seconds of end of source sound

I spontaneous (Immediate Spontaneous): while NOT oriented towards screen (within in 2 seconds of starting sound), speaker, within 5 seconds of end of source sound

D spontaneous (Delayed): while NOT oriented towards screen (within in 2 seconds of starting sound), speaker, after 5 seconds of end of source sound

Imitative: the child attempts to echo or repeat a sound previously heard/made by computer or human Must match 50% of phonemes of ATTEMPTED target OR same number of syllables as whole. Target resets after each new non-child sound. Cannot imitate self OR echoed sound

Di⊠ered imitation: time frame for Differed Imitation ends after a new sound is made by speaker, or researcher

Immediate imitation: within 5 seconds from source

PASS 4: *Drag Position for fast skim*

*Time In Chair: marking times during the video when the child is seated in the chair. This includes having the child's butt on the chair, 2 legs on the chair (ie. sitting cross-legged, sitting on feet) or otherwise in the chair in a manner generally deemed "sitting". Not included is one leg on chair, one leg standing.

Appendix B

ACES Appendix Chapter

B.1 Write-It Do-It: Potential Applications & Measures

While we discuss many theories of communication and linguistics, applying all of them to ACES output is well beyond the scope of this project. Rather, we present and describe this prior work as motivation for the need of a platform like ACES to study Aphasic language changes.

B.1.1 Communication Accommodation Theory (CAT)

The Communication Accommodation Theory (CAT) was developed by Howard Giles (dissertation in 1971(Giles, 1971), published in 1973(Giles et al., 1973)). In short, CAT attempts to explain the rationale for how and why people's language change during conversations. This could be changes (most notably convergence and divergence) in content, phrasing, vocabulary, and other speech patterns. CAT is a well accepted and well-researched theory that continues to be studied. Further, CAT is considered a major socio-psychological theory of language and social interaction (Tracy and Haspel, 2004). Some current avenues of research look across cultures for how their language changes and evolves during interactions(S'hiri, 2002), how age affects interaction in business settings (McCann and Giles, 2006), public interaction with police offers (Giles et al., 2006) and discourse over IM (Scissors et al., 2009). Using CAT allows us to study and understand these interactions in a truly significant way. Aphasia and ACES would be well situated in this literature, and well justified as an avenue of research.

B.1.1.1 Measuring CAT

A commonly used method to study CAT is through the output of Linguistic Inquiry and Word Count (LIWC). LIWC (Pennebaker et al., 2007) is a program that uses a bag-of-words approach to measure

content (e.g., judgments, personality, topics, etc.) as well as the more low level elements of language (e.g., punctuation, phrasing, etc). Overall LIWC is a hand/hard coded analysis of language across numerous dimensions. LIWC processes one, or a cohort of text files, and outputs the results in each of these dimensions, which include general categories (word count, words per sentence, etc.), linguistic categories (percentage of words in the text which are nouns, verbs, etc.), psychological constructs (e.g., affect, cognition), personal concern categories (e.g., work, home), paralinguistic categories (e.g., assents such as "mmhmm"), and punctuation categories (periods, commas, etc.).

Interesting findings from the first set of ACES logs have already been uncovered and published (Danilevsky et al., 2011). We have used some IR and Data Mining techniques to uncover linguistic changes in the conversations. We also have some initial support for alignment theory, and adaptation theory in our initial log set. Should we uncover these and other patterns, we can clearly show that by using ACES, users follow patterns of real individuals with Aphasia when having conversations; thus showing a real world impact and learning for typical individuals. Given the financial cost of using LIWC and the lack of computational sophistication, researchers may be able to add leverage (and potentially contribute) Data Mining and Information Retrieval Techniques.

B.1.2 Interactive Alignment Account or Alignment Theory

Alignment theory is a newer theory on communication that states that when two people have a conversation, their language will change and align (become similar) on many levels (e.g., verb selection, noun selection, sentence structure, etc.) and align situation model. The theory states that when languages align there is improved production and comprehension in dialogue (Pickering and Garrod, 2006). It can be thought of as a method of cooperation that occurs to allow the sentence itself to succeed. Pickering and Garrod specifically state alignment occurs on a semantic, syntactic, lexical, phonological and phonetic level. Unlike CAT which is a far broader examination of changes in linguistics and semantics, Alignment Theory specifically focuses on people using the same words to describe the same object, concept or situation. in a specific conversation. What make alignment theory so interesting, and relevant to ACES and aphasia, is the effect of “atypical language” has on the “propensity for alignment”, as well as HCI. Picker and Garrod present that it is an open question as to how atypical language (e.g., aphasia) and HCI (ACES) effects alignment.

B.2 Write-It Do-It: Question and Continuers Lists

B.2.1 Questions

The following is the list of "Questions" that were identified (by beginning a sentence) and counted through the question analysis (in combination with messages containing a question mark).

- what
- when
- where
- which
- who
- whom
- whose
- why
- how
- is
- did

We also counted the sub-type of "how" questions for release with our data set. However, the sub question analysis was not conducted as part of this dissertation.

- how far
- how long
- how many
- how much
- how old
- how should
- how can
- how could
- how is

B.2.2 Continuers

The following is the list of "Continuers" that were identified and counted through the pragmatic analysis.

- got it
- did not get it
- didn't get it
- k
- ok
- yep
- sure

Imperative Continuers (e.g. "cut" or "place") were considered, however, identifying when these terms were used as Continuers versus questions or other uses would have been far more difficult and not as easy to identify because they would not be able to be matched with regular expressions. Further this list of imperatives would have been highly variable, and require a full examination of the entire corpus ensuring that no terms were missed. Examining imperatives is important future work.

B.3 Write-It Do-It: Conversation Quality Tables

B.3.1 Write-It Do-It: Summary Statistics - Conversation Quality Measures

	Mean (sd)	Median [IQ Range]
Obj Score	0.32 (0.14)	0.31 [0.23 , 0.43]
Doer ICSI	5.48 (1.25)	6.33 [4.25 , 6.33]
Writer ICSI	5.15 (1.41)	5.33 [4.50 , 6.00]
Doer ICR	3.74 (1.17)	3.55 [2.85 , 4.55]
Writer ICR	3.79 (1.11)	3.70 [3.10 , 4.35]

Table B.1: *Summary Statistics – Conversation Quality – Control Cohort*

	Mean (sd)	Median [IQ Range]
Obj Score	0.17 (0.06)	0.18 [0.14 , 0.22]
Doer ICSI	4.64 (1.31)	4.58 [3.83 , 4.58]
Writer ICSI	4.53 (1.22)	4.42 [3.58 , 5.50]
Doer ICR	4.79 (1.15)	4.95 [4.00 , 5.45]
Writer ICR	4.87 (1.20)	4.70 [3.95 , 5.90]

Table B.2: *Summary Statistics – Conversation Quality – Aphasic Writer Cohort*

	Mean (sd)	Median [IQ Range]
Obj Score	0.31 (0.18)	0.32 [0.15 , 0.39]
Doer ICSI	4.91 (1.25)	5.25 [4.00 , 5.92]
Writer ICSI	4.99 (1.07)	5.00 [4.33 , 5.83]
Doer ICR	4.13 (1.14)	4.00 [3.45 , 4.95]
Writer ICR	4.26 (0.96)	4.25 [3.50 , 4.80]

Table B.3: *Summary Statistics – Conversation Quality – Aphasic Doer Cohort*

B.3.2 Write-It Do-It: Comparative Statistics of Conversation Quality between Cohorts

	Control vs. Aphasic Writer	Control vs. Aphasic Doer	Aphasic Writer vs. Aphasic Doer
Obj Score	< 0.01	0.84	< 0.01
Doer ICSI	0.01	0.06	0.38
Writer ICSI	0.05	0.62	0.10
Doer ICR	< 0.01	0.18	0.02
Writer ICR	< 0.01	0.07	0.02

Table B.4: *Comparative Statistics Between the Three Cohorts – Conversation Quality*
All Statistics are p-value from GEE

B.3.3 Write-It Do-It: Comparative Statistics of Conversation Quality between Writers and Doers

	Control Cohort	Aphasic Writer Cohort	Aphasic Doer Cohort
ICSI	0.31	0.74	0.77
ICR	0.87	0.80	0.61

Table B.5: *Comparative Statistics Between Writers and Doers within Each Cohort – Conversation Quality*
All Statistics are p-value from GEE comparing Writer's Scores to Doer's Scores

B.3.4 Write-It Do-It: GEE Correlations between Measures of Conversation Quality

	Objective		Writer ICSI		Writer ICR		Doer ICSI		Doer ICR	
	coef	*p-value*	*coef*	*p-value*	*coef*	*p-value*	*coef*	*p-value*	*coef*	*p-value*
Objective	1.00	0.00	—	—	—	—	—	—	—	—
Writer ICSI	0.01	0.59	1.00	0.00	—	—	—	—	—	—
Writer ICR	-0.28	0.20	-0.75	< 0.01	1.00	0.00	—	—	—	—
Doer ICSI	0.02	0.43	0.25	0.19	-0.09	0.55	1.00	0.00	—	—
Doer ICR	-0.01	0.66	-0.30	0.15	0.11	0.53	-0.89	< 0.01	1.00	0.00

Table B.6: *GEE Correlations – Conversation Quality - Control Cohort*
Results represent Coefficient and P-Values from GEE

	Objective		Writer ICSI		Writer ICR		Doer ICSI		Doer ICR	
	coef	*p-value*	*coef*	*p-value*	*coef*	*p-value*	*coef*	*p-value*	*coef*	*p-value*
Objective	1.00	0.00	—	—	—	—	—	—	—	—
Writer ICSI	0.01	0.32	1.00	0.00	—	—	—	—	—	—
Writer ICR	-0.02	0.03	-0.75	< 0.01	1.00	0.00	—	—	—	—
Doer ICSI	-0.00	0.91	0.10	0.56	-0.08	0.60	1.00	0.00	—	—
Doer ICR	-0.00	0.76	-0.14	0.46	0.15	0.40	-0.86	< 0.01	1.00	0.00

Table B.7: *GEE Correlations – Conversation Quality - Aphasic Writer Cohort*
Results represent Coefficient and P-Values from GEE

	Objective		Writer ICSI		Writer ICR		Doer ICSI		Doer ICR	
	coef	*p-value*	*coef*	*p-value*	*coef*	*p-value*	*coef*	*p-value*	*coef*	*p-value*
Objective	1.00	0.00	—	—	—	—	—	—	—	—
Writer ICSI	-0.01	0.85	1.00	0.00	—	—	—	—	—	—
Writer ICR	0.06	0.08	-0.44	0.02	1.00	0.00	—	—	—	—
Doer ICSI	0.06	0.01	-0.23	0.11	0.26	0.04	1.00	0.00	—	—
Doer ICR	-0.02	0.63	0.39	0.01	-0.10	0.48	-0.79	< 0.01	1.00	0.00

Table B.8: *GEE Correlations – Conversation Quality - Aphasic Doer Cohort*
Results represent Coefficient and P-Values from GEE

	Objective R^2	Writer ICSI R^2	Writer ICR R^2	Doer ICSI R^2	Doer ICR R^2
Objective	—	—	—	—	—
Writer ICSI	—	—	—	—	—
Writer ICR	—	-0.60	—	—	—
Doer ICSI	—	—	—	—	—
Doer ICR	—	—	—	-0.84	—

Table B.9: *Pearson's Correlations R^2 for Significant Correlations – Conversation Quality - Control Cohort*
Results represent Coefficient from Pearson's. It should be noted that Pearson's does not account for the correlated nature of the data.

	Objective R^2	Writer ICSI R^2	Writer ICR R^2	Doer ICSI R^2	Doer ICR R^2
Objective	—	—	—	—	—
Writer ICSI	—	—	—	—	—
Writer ICR	-0.36	0.74	—	—	—
Doer ICSI	—	—	—	—	—
Doer ICR	—	—	—	-0.75	—

Table B.10: *Pearson's Correlations R^2 for Significant Correlations – Conversation Quality - Aphasic Writer Cohort*
Results represent Coefficient from Pearson's. It should be noted that Pearson's does not account for the correlated nature of the data.

	Objective R^2	Writer ICSI R^2	Writer ICR R^2	Doer ICSI R^2	Doer ICR R^2
Objective	—	—	—	—	—
Writer ICSI	—	—	—	—	—
Writer ICR	—	-0.39	—	—	—
Doer ICSI	0.44	—	0.34	—	—
Doer ICR	—	0.41	—	-0.72	—

Table B.11: *Pearson's Correlations R^2 for Significant Correlations – Conversation Quality - Aphasic Doer Cohort*

Results represent Coefficient from Pearson's. It should be noted that Pearson's does not account for the correlated nature of the data.

B.3.5 Write-It Do-It: Correlation Statistics of Writers Conversation Quality and Doers Conversation Quality

	Control Cohort	Aphasic Writer Cohort	Aphasic Doer Cohort
ICSI	0.19	0.56	0.11
ICR	0.53	0.40	0.48

Table B.12: *Correlation Statistics Between Writers and Doers within Each Cohort – Conversation Quality*
All Statistics are p-value from GEE comparing Writer's Scores to Doer's Scores

B.4 Write-It Do-It: Linguistic (Raw Value/Count) Tables

B.4.1 Write-It Do-It: Summary Statistics - Writer's Linguistic Measure Value

	Mean (sd)	Median [IQ Range]
Writer's Lines	84.53 (40.63)	80.50 [55.00 , 122.50]
Writer's Function Words	342.16 (79.69)	336.00 [294.50 , 390.50]
Writer's Content Words	455.43 (97.74)	444 [396 , 519.50]
Writer's Words	797.60 (173.78)	763.00 [696.00 , 898.50]
Writer's Adjectives	90.03 (23.13)	85 [73.50 , 100.50]
Writer's Verbs	104.75 (28.48)	104.00 [86.00 , 121.50]
Writer's Nouns	221.97 (50.87)	217.50 [187.00 , 260.00]
Writer's Adverbs	38.69 (15.25)	37.00 [27.00 , 48.50]
Writer's Punctuation	51.78 (24.68)	49.00 [36.00 , 61.50]
Writer's Questions	5.38 (5.28)	3.00 [2.00 , 8.00]
Writer's Continuers	5.16 (5.20)	4.00 [2.00 , 7.50]

Table B.13: *Summary Statistics for the Control Group – Writer's Values*

	Mean (sd)	Median [IQ Range]
Writer's Lines	99.06 (54.81)	89.50 [57.00 , 115.00]
Writer's Function Words	245.25 (107.16)	240.00 [168.00 , 308.00]
Writer's Content Words	356.94 (144.91)	318.50 [226.00 , 499.00]
Writer's Words	602.19 (243.53)	565.00 [402.50 , 763.00]
Writer's Adjectives	76.03 (33.52)	66.00 [50.50 , 91.50]
Writer's Verbs	70.56 (36.27)	63.00 [43.5 , 91.5]
Writer's Nouns	186.69 (71.67)	173.00 [121.50 , 254.00]
Writer's Adverbs	23.66 (13.82)	22.00 [12.50 , 33.50]
Writer's Punctuation	41.84 (31.73)	37.00 [15.00 , 70.00]
Writer's Questions	5.19 (6.78)	3.00 [1.00 , 7.00]
Writer's Continuers	3.72 (5.28)	2.50 [0.00 , 5.00]

Table B.14: *Summary Statistics for the Aphasic Writer Group – Writer's Values*

	Mean (sd)	Median [IQ Range]
Writer's Lines	81.63 (40.90)	77.50 [53.00 , 95.50]
Writer's Function Words	335.56 (116.84)	312.50 [258.00 , 419.50]
Writer's Content Words	436.31 (134.52)	430.50 [352.50 , 515.50]
Writer's Words	771.88 (249.80)	735.00 [620.50 , 925.50]
Writer's Adjectives	84.72 (26.27)	82.00 [69.00 , 102.00]
Writer's Verbs	103.69 (38.70)	97.00 [77.00 , 125.50]
Writer's Nouns	208.84 (62.37)	198.50 [174.00 , 257.00]
Writer's Adverbs	39.06 (19.03)	39.00 [25.00 , 52.50]
Writer's Punctuation	36.25 (20.05)	32.00 [22.00 , 47.50]
Writer's Questions	6.91 (5.61)	6.50 [3.50 , 8.00]
Writer's Continuers	5.72 (6.58)	3.00 [1.00 , 7.00]

Table B.15: *Summary Statistics for the Aphasic Doer Group – Writer's Values*

B.4.2 Write-It Do-It: Summary Statistics - Doer's Linguistic Measure Value

	Mean (sd)	Median [IQ Range]
Doer's Lines	57.72 (27.29)	49.00 [35.50 , 74.50]
Doer's Function Words	112.13 (45.41)	112.50 [83.00 , 142.50]
Doer's Content Words	170.75 (60.01)	166.50 [130.00 , 219.00]
Doer's Words	282.88 (103.72)	279.50 [214.00 , 353.50]
Doer's Adjectives	28.31 (11.69)	27.00 [20.50 , 36.00]
Doer's Verbs	50.38 (19.40)	50.50 [34.00 , 60.50]
Doer's Nouns	30.75 (78.00)	79.50 [58.00 , 99.50]
Doer's Adverbs	14.06 (6.68)	14.00 [10.00 , 17.00]
Doer's Punctuation	36.03 (15.21)	36.50 [24.50 , 48.00]
Doer's Questions	16.96 (8.25)	15.00 [10.00 , 21.00]
Doer's Continuers	17.06 (11.74)	18.50 [6.5 , 24.00]

Table B.16: *Summary Statistics for the Control Group – Doer's Values*

	Mean (sd)	Median [IQ Range]
Doer's Lines	68.47 (41.97)	54.50 [38.50 , 84.00]
Doer's Function Words	147.06 (77.41)	146.00 [98.00 , 180.50]
Doer's Content Words	227.75 (119.44)	212.00 [157.00 , 272.50]
Doer's Words	376.81 (192.56)	358.00 [255.00 , 468.50]
Doer's Adjectives	43.69 (25.97)	38.50 [27.50 , 57.00]
Doer's Verbs	61.28 (34.04)	59.00 [37.50 , 78.00]
Doer's Nouns	107.44 (57.74)	99.00 [65.50 , 129.50]
Doer's Adverbs	15.34 (10.29)	13.00 [8.00 , 20.50]
Doer's Punctuation	55.12 (30.49)	55.00 [32.50 , 71.00]
Doer's Questions	31.44 (19.60)	27.00 [17.50 , 37.00]
Doer's Continuers	16.25 (12.33)	13.00 [6.50 , 27.50]

Table B.17: *Summary Statistics for the Aphasic Writer Group – Doer's Values*

	Mean (sd)			Median [IQ Range]			
Doer's Lines	56.25	(22.78	)	51.50	[38.50	, 70.50	]
Doer's Function Words	90.00	(53.02	)	84.50	[55.00	, 113.50	]
Doer's Content Words	138.16	(73.55	)	114.50	[88.50	, 184.00	]
Doer's Words	228.16	(124.67	)	199.00	[145.50	, 286.50	]
Doer's Adjectives	25.34	(18.93	)	22.00	[9.00	, 39.50	]
Doer's Verbs	41.59	(21.72	)	41.50	[26.00	, 52.50	]
Doer's Nouns	61.19	(36.00	)	52.00	[38.50	, 82.50	]
Doer's Adverbs	10.03	(6.94	)	9.00	[4.50	, 14.50	]
Doer's Punctuation	25.09	(16.81	)	24.50	[11.50	, 33.50	]
Doer's Questions	15.34	(8.68	)	16.00	[9.50	, 19.50	]
Doer's Continuers	14.47	(11.02	)	13.00	[4.50	, 19.00	]

Table B.18: *Summary Statistics for the Aphasic Doer Group – Doer's Values*

B.4.3 Write-It Do-It: Comparative Statistics of Writer's and Doer's Linguistic Measure Values

	Control vs. Aphasic Writer	Control vs. Aphasic Doer	Aphasic Writer vs. Aphasic Doer
Writer's Lines	0.22	0.77	0.14
Writer's Function Words	< 0.01	0.79	< 0.01
Writer's Content Words	< 0.01	0.51	0.02
Writer's Words	< 0.01	0.63	< 0.01
Writer's Adjectives	0.05	0.38	0.24
Writer's Verbs	< 0.01	0.90	< 0.01
Writer's Nouns	0.02	0.35	0.18
Writer's Adverbs	< 0.01	0.93	< 0.01
Writer's Punctuation	0.16	0.01	0.39
Writer's Questions	0.90	0.25	0.26
Writer's Continuers	0.27	0.70	0.17

Table B.19: *Comparative Statistics Between the Three Cohorts – Writer's Linguistic Measures (as raw occurrence values)*

All Statistics are p-value from GEE

	Control vs. Aphasic Writer	Control vs. Aphasic Doer	Aphasic Writer vs. Aphasic Doer
Doer's Lines	0.22	0.81	0.14
Doer's Function Words	0.02	0.07	< 0.01
Doer's Content Words	0.01	0.05	< 0.01
Doer's Words	0.01	0.05	< 0.01
Doer's Adjectives	< 0.01	0.44	< 0.01
Doer's Verbs	0.11	0.08	0.01
Doer's Nouns	0.01	0.04	< 0.01
Doer's Adverbs	0.55	0.02	0.01
Doer's Punctuation	< 0.01	0.01	< 0.01
Doer's Questions	< 0.01	0.44	< 0.01
Doer's Continuers	0.78	0.35	0.54

Table B.20: *Comparative Statistics Between the Three Cohorts – Doer's Linguistic Measures (as raw occurrence values)*

All Statistics are p-value from GEE

B.4.4 Write-It Do-It: Correlation Statistics of Writers Linguistic Measure Values and Doers Linguistic Measure Values

	Control Cohort	Aphasic Writer Cohort	Aphasic Doer Cohort
Count Lines	< 0.01	< 0.01	< 0.01
Count Function Words	0.90	0.94	0.66
Count Content Words	0.95	0.15	0.99
Count All Words	0.97	0.50	0.88
Count Adjectives	0.97	0.12	0.42
Count Verbs	0.15	0.45	0.60
Count Nouns	1.00	0.07	0.75
Count Adverbs	0.93	0.97	0.71
Count Punctuation	0.24	0.62	0.29
Count Questions	0.33	0.83	0.48
Count Continuers	0.08	0.01	0.85

Table B.21: *Correlation Statistics Between Writers and Doers within Each Cohort – Linguistic Measure Values*

All Statistics are p-value from GEE comparing Writer's Count to Doer's Count all Words that are of specified typee

B.4.5 Write-It Do-It: Correlations between Writer's Linguistic Measure Valuess and Objective Conversation Quality

	Coefficient	p-value
Writer's Lines	161.5	< 0.01
Writer's Function Words	92.27	0.01
Writer's Content Words	274.46	0.02
Writer's Words	503.65	0.01
Writer's Adjectives	11.67	0.69
Writer's Verbs	90.10	< 0.01
Writer's Nouns	122.04	0.04
Writer's Adverbs	50.65	< 0.01
Writer's Punctuation	-0.20	1.00
Writer's Questions	12.31	< 0.01
Writer's Continuers	11.74	0.06

Table B.22: *Correlations - Writer's Values of Linguistic Measures With Objective Conversation Quality - Control Cohort*

GEE coefficient indicates slope of correlation
GEE p-value indicates how correlated the distributions are to each other

	Coefficient	p-value
Writer's Lines	201.03	0.19
Writer's Function Words	295.39	0.33
Writer's Content Words	552.28	0.18
Writer's Words	847.68	0.22
Writer's Adjectives	160.30	0.08
Writer's Verbs	61.67	0.55
Writer's Nouns	303.45	0.13
Writer's Adverbs	26.87	0.50
Writer's Punctuation	72.38	0.42
Writer's Questions	-8.26	0.67
Writer's Continuers	6.54	0.67

Table B.23: *Correlations - Writer's Values of Linguistic Measures With Objective Conversation Quality - Aphasic Writer Cohort*

GEE coefficient indicates slope of correlation
GEE p-value indicates how correlated the distributions are to each other

	Coefficient	p-value
Writer's Lines	113.60	< 0.01
Writer's Function Words	318.73	< 0.01
Writer's Content Words	368.66	< 0.01
Writer's Words	687.40	< 0.01
Writer's Adjectives	62.34	0.01
Writer's Verbs	92.83	0.01
Writer's Nouns	156.94	< 0.01
Writer's Adverbs	56.54	< 0.01
Writer's Punctuation	11.30	0.56
Writer's Questions	4.95	0.36
Writer's Continuers	10.23	0.10

Table B.24: *Correlations - Writer's Values of Linguistic Measures With Objective Conversation Quality - Aphasic Doer Cohort*

GEE coefficient indicates slope of correlation

GEE p-value indicates how correlated the distributions are to each other

B.4.6 Write-It Do-It: Correlations between Writer's Values of Linguistic Measures and Writer's ICSI

	Coefficient	p-value
Writer's Lines	-5.46	0.27
Writer's Function Words	-6.27	0.53
Writer's Content Words	-5.97	0.63
Writer's Words	-12.24	0.57
Writer's Adjectives	0.89	0.76
Writer's Verbs	-1.79	0.62
Writer's Nouns	-4.45	0.48
Writer's Adverbs	-0.63	0.74
Writer's Punctuation	3.23	0.29
Writer's Questions	0.22	0.74
Writer's Continuers	0.80	0.21

Table B.25: *Correlations - Writer's Values of Linguistic Measures With Writer's ICSI Conversation Quality - Control Cohort*

GEE coefficient indicates slope of correlation
GEE p-value indicates how correlated the distributions are to each other

	Coefficient	p-value
Writer's Lines	-9.58	0.22
Writer's Function Words	-5.22	0.74
Writer's Content Words	-19.02	0.36
Writer's Words	-24.24	0.49
Writer's Adjectives	-2.06	0.67
Writer's Verbs	-4.36	0.40
Writer's Nouns	-10.76	0.29
Writer's Adverbs	-1.84	0.35
Writer's Punctuation	-3.90	0.39
Writer's Questions	-2.10	0.02
Writer's Continuers	-0.82	0.28

Table B.26: *Correlations - Writer's Values of Linguistic Measures With Writer's ICSI Conversation Quality - Aphasic Writer Cohort*

GEE coefficient indicates slope of correlation
GEE p-value indicates how correlated the distributions are to each other

	Coefficient	p-value
Writer's Lines	2.34	0.73
Writer's Function Words	30.44	0.10
Writer's Content Words	29.61	9.17
Writer's Words	60.05	0.13
Writer's Adjectives	8.11	0.05
Writer's Verbs	4.14	0.51
Writer's Nouns	15.39	0.12
Writer's Adverbs	1.96	0.53
Writer's Punctuation	0.04	0.99
Writer's Questions	-1.29	0.15
Writer's Continuers	-1.98	0.05

Table B.27: *Correlations - Writer's Values of Linguistic Measures With Writer's ICSI Conversation Quality - Aphasic Doer Cohort*

GEE coefficient indicates slope of correlation

GEE p-value indicates how correlated the distributions are to each other

B.4.7 Write-It Do-It: Correlations between Writer's Values of Linguistic Measures and Doer's ICSI

	Coefficient	p-value
Writer's Lines	9.29	0.09
Writer's Function Words	31.98	< 0.01
Writer's Content Words	34.71	0.01
Writer's Words	66.69	< 0.01
Writer's Adjectives	2.07	0.53
Writer's Verbs	11.92	< 0.01
Writer's Nouns	14.26	0.04
Writer's Adverbs	6.46	< 0.01
Writer's Punctuation	1.59	0.65
Writer's Questions	0.94	0.20
Writer's Continuers	1.65	0.02

Table B.28: *Correlations - Writer's Values of Linguistic Measures With Doer's ICSI Conversation Quality - Control Cohort*

GEE coefficient indicates slope of correlation
GEE p-value indicates how correlated the distributions are to each other

	Coefficient	p-value
Writer's Lines	-6.16	0.40
Writer's Function Words	-13.21	0.35
Writer's Content Words	-21.02	0.27
Writer's Words	-34.23	0.29
Writer's Adjectives	-5.20	0.24
Writer's Verbs	-4.32	0.37
Writer's Nouns	-0.24	0.33
Writer's Adverbs	-2.26	0.21
Writer's Punctuation	1.21	0.78
Writer's Questions	-0.32	0.72
Writer's Continuers	-0.23	0.75

Table B.29: *Correlations - Writer's Values of Linguistic Measures With Doer's ICSI Conversation Quality - Aphasic Writer Cohort*

GEE coefficient indicates slope of correlation
GEE p-value indicates how correlated the distributions are to each other

	Coefficient	p-value
Writer's Lines	9.66	0.08
Writer's Function Words	33.58	0.03
Writer's Content Words	38.21	0.03
Writer's Words	71.80	0.03
Writer's Adjectives	6.17	0.08
Writer's Verbs	12.45	0.01
Writer's Nouns	14.39	0.09
Writer's Adverbs	5.21	0.04
Writer's Punctuation	-0.42	0.88
Writer's Questions	1.27	0.10
Writer's Continuers	2.04	0.02

Table B.30: *Correlations - Writer's Values of Linguistic Measures With Doer's ICSI Conversation Quality - Aphasic Doer Cohort*

GEE coefficient indicates slope of correlation

GEE p-value indicates how correlated the distributions are to each other

B.4.8 Write-It Do-It: Correlations between Doer's Linguistic Measure Valuess and Objective Conversation Quality

	Coefficient	p-value
Doer's Lines	102.09	< 0.01
Doer's Function Words	75.46	0.18
Doer's Content Words	129.51	0.07
Doer's Words	204.98	0.10
Doer's Adjectives	19.33	0.18
Doer's Verbs	9.05	0.71
Doer's Nouns	101.92	< 0.01
Doer's Adverbs	-0.78	0.93
Doer's Punctuation	40.89	0.02
Doer's Questions	34.79	< 0.01
Doer's Continuers	55.39	< 0.01

Table B.31: *Correlations - Doer's Values of Linguistic Measures With Objective Conversation Quality - Control Cohort*

GEE coefficient indicates slope of correlation
GEE p-value indicates how correlated the distributions are to each other

	Coefficient	p-value
Doer's Lines	105.99	0.38
Doer's Function Words	-66.26	0.77
Doer's Content Words	55.38	0.87
Doer's Words	-10.89	0.98
Doer's Adjectives	22.86	0.75
Doer's Verbs	-92.19	0.34
Doer's Nouns	134.05	0.42
Doer's Adverbs	-10.34	0.73
Doer's Punctuation	-72.79	0.40
Doer's Questions	-13.55	0.81
Doer's Continuers	38.56	0.27

Table B.32: *Correlations - Doer's Values of Linguistic Measures With Objective Conversation Quality - Aphasic Writer Cohort*

GEE coefficient indicates slope of correlation
GEE p-value indicates how correlated the distributions are to each other

	Coefficient	p-value
Doer's Lines	60.91	< 0.01
Doer's Function Words	83.50	0.09
Doer's Content Words	104.74	0.13
Doer's Words	188.24	0.11
Doer's Adjectives	34.77	0.05
Doer's Verbs	41.84	0.03
Doer's Nouns	29.63	0.39
Doer's Adverbs	-1.50	0.82
Doer's Punctuation	8.90	0.58
Doer's Questions	6.28	0.45
Doer's Continuers	14.72	0.16

Table B.33: *Correlations - Doer's Values of Linguistic Measures With Objective Conversation Quality - Aphasic Doer Cohort*

GEE coefficient indicates slope of correlation

GEE p-value indicates how correlated the distributions are to each other

B.4.9 Write-It Do-It: Correlations between Doer's Values of Linguistic Measures and Writer's ICSI

	Coefficient	p-value
Doer's Lines	1.67	0.63
Doer's Function Words	3.70	0.51
Doer's Content Words	-1.29	0.86
Doer's Words	2.41	0.85
Doer's Adjectives	-3.39	0.01
Doer's Verbs	0.56	0.82
Doer's Nouns	1.38	0.72
Doer's Adverbs	0.16	0.85
Doer's Punctuation	-0.88	0.65
Doer's Questions	-0.45	0.66
Doer's Continuers	2.21	0.12

Table B.34: *Correlations - Doer's Values of Linguistic Measures With Writer's ICSI Conversation Quality - Control Cohort*

GEE coefficient indicates slope of correlation
GEE p-value indicates how correlated the distributions are to each other

	Coefficient	p-value
Doer's Lines	-2.66	0.66
Doer's Function Words	3.09	0.78
Doer's Content Words	-1.32	0.94
Doer's Words	1.77	0.95
Doer's Adjectives	0.87	0.81
Doer's Verbs	-1.91	0.70
Doer's Nouns	-0.70	0.93
Doer's Adverbs	0.41	0.78
Doer's Punctuation	4.92	0.26
Doer's Questions	1.91	0.50
Doer's Continuers	1.54	0.38

Table B.35: *Correlations - Doer's Values of Linguistic Measures With Writer's ICSI Conversation Quality - Aphasic Writer Cohort*

GEE coefficient indicates slope of correlation
GEE p-value indicates how correlated the distributions are to each other

	Coefficient	p-value
Doer's Lines	-5.43	-.14
Doer's Function Words	-15.47	0.45
Doer's Content Words	-7.24	0.01
Doer's Words	-0.13	0.97
Doer's Adjectives	-3.98	0.50
Doer's Verbs	-2.09	0.05
Doer's Nouns	-4.46	0.09
Doer's Adverbs	-1.19	0.40
Doer's Punctuation	0.61	0.74
Doer's Questions	9.29	0.09
Doer's Continuers	31.98	< 0.01

Table B.36: *Correlations - Doer's Values of Linguistic Measures With Writer's ICSI Conversation Quality - Aphasic Doer Cohort*

GEE coefficient indicates slope of correlation

GEE p-value indicates how correlated the distributions are to each other

B.4.10 Write-It Do-It: Correlations between Doer's Values of Linguistic Measures and Doer's ICSI

	Coefficient	p-value
Doer's Lines	1.65	0.02
Doer's Function Words	6.70	0.07
Doer's Content Words	7.49	0.24
Doer's Words	13.76	0.34
Doer's Adjectives	-1.79	0.27
Doer's Verbs	5.09	0.05
Doer's Nouns	1.83	0.67
Doer's Adverbs	1.13	0.22
Doer's Punctuation	-0.04	0.98
Doer's Questions	0.37	0.75
Doer's Continuers	4.06	< 0.01

Table B.37: *Correlations - Doer's Values of Linguistic Measures With Doer's ICSI Conversation Quality - Control Cohort*

GEE coefficient indicates slope of correlation
GEE p-value indicates how correlated the distributions are to each other

	Coefficient	p-value
Doer's Lines	-5.24	0.35
Doer's Function Words	-2.33	0.82
Doer's Content Words	-9.26	0.56
Doer's Words	-11.60	0.65
Doer's Adjectives	0.07	0.99
Doer's Verbs	-4.98	0.27
Doer's Nouns	-2.28	0.77
Doer's Adverbs	-2.07	0.12
Doer's Punctuation	-1.30	0.75
Doer's Questions	-1.24	0.64
Doer's Continuers	-2.35	0.14

Table B.38: *Correlations - Doer's Values of Linguistic Measures With Doer's ICSI Conversation Quality - Aphasic Writer Cohort*

GEE coefficient indicates slope of correlation
GEE p-value indicates how correlated the distributions are to each other

	Coefficient	p-value
Doer's Lines	0.75	0.82
Doer's Function Words	-0.24	0.98
Doer's Content Words	0.96	0.93
Doer's Words	0.73	0.97
Doer's Adjectives	2.16	0.42
Doer's Verbs	0.68	0.83
Doer's Nouns	-1.28	0.80
Doer's Adverbs	-0.60	0.54
Doer's Punctuation	-0.30	0.90
Doer's Questions	-1.48	0.22
Doer's Continuers	-0.26	0.87

Table B.39: *Correlations - Doer's Values of Linguistic Measures With Doer's ICSI Conversation Quality - Aphasic Doer Cohort*

GEE coefficient indicates slope of correlation

GEE p-value indicates how correlated the distributions are to each other

B.4.11 Write-It Do-It: Comparative Statistics of Linguistic Measure Values between Writers and Doers

	Control Cohort	Aphasic Writer Cohort	Aphasic Doer Cohort
Count Lines	< 0.01	0.01	< 0.01
Count Function Words	< 0.01	< 0.01	< 0.01
Count Content Words	< 0.01	< 0.01	< 0.01
Count All Words	< 0.01	< 0.01	< 0.01
Count Adjectives	< 0.01	< 0.01	< 0.01
Count Verbs	< 0.01	0.28	< 0.01
Count Nouns	< 0.01	< 0.01	< 0.01
Count Adverbs	< 0.01	< 0.01	< 0.01
Count Punctuation	< 0.01	0.08	0.01
Count Questions	< 0.01	< 0.01	< 0.01
Count Continuers	< 0.01	< 0.01	< 0.01

Table B.40: *Comparative Statistics Between Writers and Doers within Each Cohort – Linguistic Measure Values*

All Statistics are p-value from GEE comparing Writer's Count to Doer's Count all Words that are of specified typee

B.5 Write-It Do-It: Linguistic (Percentage) Tables

B.5.1 Write-It Do-It: Summary Statistics - Writer's Linguistic Measure Values as Percentages

	Mean (sd)	Median [IQ Range]
% Function Words	0.43 (0.02)	0.43 [0.42 , 0.44]
% Content Words	0.57 (0.02)	0.57 [0.56 , 0.58]
% Adjectives	0.12 (0.03)	0.12 [0.09 , 0.13]
% Verbs	0.13 (0.02)	0.13 [0.12 , 0.15]
% Nouns	0.28 (0.03)	0.28 [0.26 , 0.29]
% Adverbs	0.05 (0.01)	0.05 [0.04 , 0.05]
% Questions†	0.06 (0.05)	0.05 [0.02 , 0.09]
% Continuers†	0.06 (0.6)	0.05 [0.02 , 0.08]

Table B.41: *Summary Statistics for the Control Cohort – Writer's Values as Percentages*
All values are % (in decimal form) of all Words that are of a specific type
† is percentage (in decimal form) of type per line

	Mean (sd)	Median [IQ Range]
% Function Words	0.40 (0.06)	0.42 [0.38 , 0.45]
% Content Words	0.60 (0.06)	0.58 [0.55 , 0.62]
% Adjectives	0.13 (0.03)	0.12 [0.11 , 0.15]
% Verbs	0.11 (0.02)	0.11 [0.10 , 0.13]
% Nouns	0.32 (0.48)	0.31 [0.29 , 0.34]
% Adverbs	0.04 (0.01)	0.04 [0.03 , 0.05]
% Questions†	0.05 (0.05)	0.04 [0.02 , 0.07]
% Continuers†	0.04 (0.04)	0.03 [0.00 , 0.05]

Table B.42: *Summary Statistics for the Aphasic Writer Cohort – Writer's Values as Percentages*
All values are % (in decimal form) of all Words that are of a specific type
† is percentage (in decimal form) of type per line

	Mean (sd)	Median [IQ Range]
% Function Words	0.43 (0.03)	0.43 [0.42 , 0.45]
% Content Words	0.57 (0.03)	0.57 [0.55 , 0.58]
% Adjectives	0.11 (0.02)	0.11 [0.10 , 0.12]
% Verbs	0.13 (0.02)	0.14 [0.12 , 0.15]
% Nouns	0.27 (0.03)	0.28 [0.26 , 0.29]
% Adverbs	0.05 (0.01)	0.05 [0.04 , 0.06]
% Questions†	0.09 (0.06)	0.07 [0.04 , 0.13]
% Continuers†	0.07 (0.07)	0.05 [0.02 , 0.10]

Table B.43: *Summary Statistics for the Aphasic Doer Cohort – Writer's Values as Percentages*
All values are % (in decimal form) of all Words that are of a specific type
† is percentage (in decimal form) of type per line

B.5.2 Write-It Do-It: Summary Statistics - Doer's Linguistic Measure Values as Percentages

	Mean (sd)	Median [IQ Range]
% Function Words	0.39 (0.04)	0.39 [0.36 , 0.42]
% Content Words	0.61 (0.04)	0.61 [0.58 , 0.64]
% Adjectives	0.10 (0.04)	0.10 [0.08 , 0.12]
% Verbs	0.18 (0.03)	0.18 [0.16 , 0.21]
% Nouns	0.28 (0.04)	0.27 [0.25 , 0.30]
% Adverbs	0.05 (0.02)	0.05 [0.04 , 0.06]
% Questions†	0.31 (0.11)	0.29 [0.21 , 0.39]
% Continuers†	0.29 (0.20)	0.31 [0.12 , 0.45]

Table B.44: *Summary Statistics for the Control Cohort – Writer's Values as Percentages*
All values are % (in decimal form) of all Words that are of a specific type
† is percentage (in decimal form) of type per line

	Mean (sd)	Median [IQ Range]
% Function Words	0.40 (0.04)	0.40 [0.38 , 0.42]
% Content Words	0.60 (0.04)	0.60 [0.58 , 0.62]
% Adjectives	0.11 (0.03)	0.11 [0.09 , 0.13]
% Verbs	0.16 (0.03)	0.17 [0.14 , 0.19]
% Nouns	0.29 (0.05)	0.28 [0.26 , 0.31]
% Adverbs	0.04 (0.01)	0.04 [0.03 , 0.05]
% Questions†	0.49 (0.19)	0.47 [0.38 , 0.63]
% Continuers†	0.24 (0.14)	0.27 [0.16 , 0.33]

Table B.45: *Summary Statistics for the Aphasic Writer Cohort – Writer's Values as Percentages*
All values are % (in decimal form) of all Words that are of a specific type
† is percentage (in decimal form) of type per line

	Mean (sd)	Median [IQ Range]
% Function Words	0.38 (0.06)	0.39 [0.37 , 0.42]
% Content Words	0.62 (0.06)	0.61 [0.58 , 0.63]
% Adjectives	0.10 (0.06)	0.09 [0.06 , 0.15]
% Verbs	0.20 (0.11)	0.18 [0.16 , 0.21]
% Nouns	0.27 (0.07)	0.27 [0.22 , 0.33]
% Adverbs	0.04 (0.02)	0.04 [0.03 , 0.06]
% Questions†	0.28 (0.13)	0.28 [0.18 , 0.36]
% Continuers†	0.28 (0.21)	0.21 [0.11 , 0.49]

Table B.46: *Summary Statistics for the Aphasic Doer Cohort – Writer's Values as Percentages*
All values are % (in decimal form) of all Words that are of a specific type
† is percentage (in decimal form) of type per line

B.5.3 Write-It Do-It: Comparative Statistics of Linguistic Measure Values as Percentages Across Cohorts

	Control vs. Aphasic Writer	Control vs. Aphasic Doer	Aphasic Writer vs. Aphasic Doer
% Function Words	0.03	0.69	0.02
% Content Words	0.03	0.69	0.02
% Adjectives	0.07	0.73	0.03
% Verbs	< 0.01	0.67	< 0.01
% Nouns	< 0.01	0.57	< 0.01
% Adverbs	< 0.01	0.69	< 0.01
% Questions†	0.39	0.08	0.01
% Continuers†	0.03	0.65	0.01

Table B.47: *Comparative Statistics Between the Three Cohorts – Writer's Value of Linguistic Measures as Percentages*
All Statistics are p-value from GEE analysis of % (in decimal form) of all Words that are of specified type
† is percentage (in decimal form) of type per line

	Control vs. Aphasic Writer	Control vs. Aphasic Doer	Aphasic Writer vs. Aphasic Doer
% Function Words	0.53	0.54	0.28
% Content Words	0.53	0.54	0.28
% Adjectives	0.32	1.00	0.47
% Verbs	0.03	0.30	0.06
% Nouns	0.32	0.71	0.27
% Adverbs	0.01	0.12	0.55
% Questions†	< 0.01	0.27	< 0.01
% Continuers†	0.24	0.78	0.43

Table B.48: *Comparative Statistics Between the Three Cohorts – Doer's Value of Linguistic Measures as Percentages*
All Statistics are p-value from GEE analysis of % (in decimal form) of all Words that are of specified type
† is percentage (in decimal form) of type per line

B.5.4 Write-It Do-It: Comparative Statistics of Linguistic Measure Values as Percentages between Writers and Doers

	Control Cohort	Aphasic Writer Cohort	Aphasic Doer Cohort
% Function Words	< 0.01	0.69	< 0.01
% Content Words	< 0.01	0.69	< 0.01
% Adjectives	0.15	0.05	0.41
% Verbs	< 0.01	< 0.01	< 0.01
% Nouns	0.80	0.02	0.77
% Adverbs	0.55	0.62	0.14
% Questions†	< 0.01	< 0.01	< 0.01
% Continuers†	< 0.01	< 0.01	< 0.01

Table B.49: *Comparative Statistics Between Writers and Doers within Each Cohort – Linguistic Measure Values as Percentages*
All Statistics are p-value from GEE comparing Writer's Scores to Doer's Scores of % (in decimal form) of all Words that are of specified type
† is percentage (in decimal form) of type per line

B.5.5 Write-It Do-It: Correlations between Writer's Linguistic Measure Values as Percentages and Objective Conversation Quality

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Writer's % Function Words	0.02	0.58	–
Writer's % Content Words	-0.02	0.58	–
Writer's % Adjectives	-0.06	0.07	–
Writer's % Verbs	0.03	0.23	–
Writer's % Nouns	-0.02	0.53	–
Writer's % Adverbs	0.03	0.01	0.40
Writer's % Questions	0.07	0.33	–
Writer's % Continuers	0.05	0.48	–

Table B.50: *Correlations - Writer's Values of Linguistic Measures as Percentages With Objective Conversation Quality - Control Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Writer's % Function Words	0.04	0.85	–
Writer's % Content Words	-0.04	0.85	–
Writer's % Adjectives	0.03	0.73	–
Writer's % Verbs	-0.04	0.46	–
Writer's % Nouns	-0.02	0.88	–
Writer's % Adverbs	-0.005	0.89	–
Writer's % Questions	-0.20	0.19	–
Writer's % Continuers	0.01	0.94	–

Table B.51: *Correlations - Writer's Values of Linguistic Measures as Percentages With Objective Conversation Quality - Aphasic Writer Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Writer's % Function Words	0.04	0.07	–
Writer's % Content Words	-0.04	0.07	–
Writer's % Adjectives	-0.03	0.28	–
Writer's % Verbs	0.01	0.71	–
Writer's % Nouns	-0.06	0.02	-0.38
Writer's % Adverbs	0.03	0.02	0.39
Writer's % Questions	-0.05	0.43	–
Writer's % Continuers	0.05	0.47	–

Table B.52: *Correlations - Writer's Values of Linguistic Measures as Percentages With Objective Conversation Quality - Aphasic Doer Cohort*

When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

B.5.6 Write-It Do-It: Correlations between Writer's Values of Linguistic Measures as Percentages and Writer's ICR

	GEE		Pearson's Correlation
	Coefficient	**p-value**	**R^2**
Writer's % Function Words	<0.01	0.70	–
Writer's % Content Words	<-0.01	0.70	–
Writer's % Adjectives	<-0.01	0.44	–
Writer's % Verbs	<-0.01	0.36	–
Writer's % Nouns	<0.01	0.37	–
Writer's % Adverbs	<0.01	0.45	–
Writer's % Questions	<0.01	1.00	–
Writer's % Continuers	-0.01	0.12	–

Table B.53: *Correlations - Writer's Values of Linguistic Measures as Percentages With Writer's ICR Conversation Quality - Control Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	**p-value**	**R^2**
Writer's % Function Words	-0.01	0.54	–
Writer's % Content Words	0.01	0.54	–
Writer's % Adjectives	0.01	0.24	–
Writer's % Verbs	<-0.01	0.83	–
Writer's % Nouns	<0.01	0.96	–
Writer's % Adverbs	<0.01	0.82	–
Writer's % Questions	0.01	0.51	–
Writer's % Continuers	<-0.01	0.67	–

Table B.54: *Correlations - Writer's Values of Linguistic Measures as Percentages With Writer's ICR Conversation Quality - Aphasic Writer Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	**p-value**	**R^2**
Writer's % Function Words	<-0.01	0.28	–
Writer's % Content Words	0.01	0.28	–
Writer's % Adjectives	<-0.01	0.77	–
Writer's % Verbs	0.01	0.08	–
Writer's % Nouns	<-0.01	0.80	–
Writer's % Adverbs	<-0.01	0.73	–
Writer's % Questions	0.01	0.38	–
Writer's % Continuers	<0.01	0.85	–

Table B.55: *Correlations - Writer's Values of Linguistic Measures as Percentages With Writer's ICR Conversation Quality - Aphasic Doer Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

B.5.7 Write-It Do-It: Correlations between Writer's Values of Linguistic Measures as Percentages and Doer's ICR

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Writer's % Function Words	<-0.01	0.27	–
Writer's % Content Words	<0.01	0.27	–
Writer's % Adjectives	<0.01	0.32	–
Writer's % Verbs	<-0.01	0.34	–
Writer's % Nouns	0.01	0.11	–
Writer's % Adverbs	<-0.01	0.05	-0.33
Writer's % Questions	<-0.01	0.69	–
Writer's % Continuers	-0.02	0.06	–

Table B.56: *Correlations - Writer's Values of Linguistic Measures as Percentages With Doer's ICR Conversation Quality - Control Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Writer's % Function Words	<0.01	0.72	–
Writer's % Content Words	<-0.01	0.72	–
Writer's % Adjectives	<-0.01	0.66	–
Writer's % Verbs	<0.01	0.14	–
Writer's % Nouns	-0.01	0.20	–
Writer's % Adverbs	<0.01	0.04	0.34
Writer's % Questions	<0.01	0.58	–
Writer's % Continuers	0.01	0.24	–

Table B.57: *Correlations - Writer's Values of Linguistic Measures as Percentages With Doer's ICR Conversation Quality - Aphasic Writer Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Writer's % Function Words	-0.01	1.00	–
Writer's % Content Words	0.01	1.00	–
Writer's % Adjectives	<0.01	0.70	–
Writer's % Verbs	-0.01	0.13	–
Writer's % Nouns	0.01	0.11	–
Writer's % Adverbs	<-0.01	0.17	–
Writer's % Questions	-0.01	0.18	–
Writer's % Continuers	-0.02	0.08	–

Table B.58: *Correlations - Writer's Values of Linguistic Measures as Percentages With Doer's ICR Conversation Quality - Aphasic Doer Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

B.5.8 Write-It Do-It: Correlations between Writer's Values of Linguistic Measures as Percentages and Writer's ICSI

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Writer's % Function Words	<-0.01	0.64	–
Writer's % Content Words	<0.01	0.64	–
Writer's % Adjectives	<0.01	0.31	–
Writer's % Verbs	<-0.01	0.88	–
Writer's % Nouns	<-0.01	0.62	–
Writer's % Adverbs	<-0.01	0.92	–
Writer's % Questions	<0.01	0.82	–
Writer's % Continuers	0.01	0.06	–

Table B.59: *Correlations - Writer's Values of Linguistic Measures as Percentages With Writer's ICSI Conversation Quality - Control Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Writer's % Function Words	0.01	0.18	–
Writer's % Content Words	-0.01	0.18	–
Writer's % Adjectives	<-0.01	0.68	–
Writer's % Verbs	<-0.01	0.90	–
Writer's % Nouns	-0.01	0.18	–
Writer's % Adverbs	<-0.01	0.60	–
Writer's % Questions	-0.02	0.03	-0.35
Writer's % Continuers	<-0.01c	0.69	–

Table B.60: *Correlations - Writer's Values of Linguistic Measures as Percentages With Writer's ICSI Conversation Quality - Aphasic Writer Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	**p-value**	**R^2**
Writer's % Function Words	0.01	0.03	0.36
Writer's % Content Words	-0.01	0.03	-0.36
Writer's % Adjectives	<-0.01	0.92	–
Writer's % Verbs	-0.01	0.12	–
Writer's % Nouns	<-0.01	0.74	–
Writer's % Adverbs	<-0.01	0.50	–
Writer's % Questions	-0.02	0.08	–
Writer's % Continuers	-0.03	0.01	-0.41

Table B.61: *Correlations - Writer's Values of Linguistic Measures as Percentages With Writer's ICSI Conversation Quality - Aphasic Doer Cohort*

When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

B.5.9 Write-It Do-It: Correlations between Writer's Values of Linguistic Measures as Percentages and Doer's ICSI

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Writer's % Function Words	<0.01	0.17	–
Writer's % Content Words	<-0.01	0.17	–
Writer's % Adjectives	-0.01	0.05	-0.33
Writer's % Verbs	<0.01	0.12	–
Writer's % Nouns	-0.01	0.10	–
Writer's % Adverbs	<0.01	< 0.01	0.47
Writer's % Questions	<0.01	0.78	–
Writer's % Continuers	0.01	0.06	–

Table B.62: *Correlations - Writer's Values of Linguistic Measures as Percentages With Doer's ICSI Conversation Quality - Control Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Writer's % Function Words	<0.01	0.62	–
Writer's % Content Words	<-0.01	0.62	–
Writer's % Adjectives	<-0.01	0.39	–
Writer's % Verbs	<0.01	0.69	–
Writer's % Nouns	<-0.01	0.90	–
Writer's % Adverbs	<-0.01	0.66	–
Writer's % Questions	<0.01	0.69	–
Writer's % Continuers	<0.01	0.88	–

Table B.63: *Correlations - Writer's Values of Linguistic Measures as Percentages With Doer's ICSI Conversation Quality - Aphasic Writer Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Writer's % Function Words	<0.01	0.42	–
Writer's % Content Words	<-0.01	0.42	–
Writer's % Adjectives	<-0.01	0.55	–
Writer's % Verbs	<0.01	0.27	–
Writer's % Nouns	<-0.01	0.08	–
Writer's % Adverbs	<0.01	0.26	–
Writer's % Questions	0.01	0.31	–
Writer's % Continuers	0.02	0.04	0.35

Table B.64: *Correlations - Writer's Values of Linguistic Measures as Percentages With Doer's ICSI Conversation Quality - Aphasic Doer Cohort*

When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

B.5.10 Write-It Do-It: Correlations between Doer's Linguistic Measure Values as Percentages and Objective Conversation Quality

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Doer's % Function Words	0.01	0.80	–
Doer's % Content Words	-0.01	0.80	–
Doer's % Adjectives	-0.04	0.39	–
Doer's % Verbs	-0.08	0.04	-0.34
Doer's % Nouns	0.13	< 0.01	0.42
Doer's % Adverbs	-0.02	0.25	–
Doer's % Questions	0.03	0.83	–
Doer's % Continuers	0.61	0.01	0.43

Table B.65: *Correlations - Doer's Values of Linguistic Measures as Percentages With Objective Conversation Quality - Control Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Doer's % Function Words	-0.17	0.16	–
Doer's % Content Words	0.17	0.16	–
Doer's % Adjectives	0.07	0.46	–
Doer's % Verbs	-0.19	0.03	-0.37
Doer's % Nouns	0.30	0.05	0.33
Doer's % Adverbs	-0.004	0.91	–
Doer's % Questions	-0.78	0.15	–
Doer's % Continuers	0.36	0.39	–

Table B.66: *Correlations - Doer's Values of Linguistic Measures as Percentages With Objective Conversation Quality - Aphasic Doer Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	**p-value**	**R^2**
Doer's % Function Words	0.07	0.23	–
Doer's % Content Words	-0.07	0.23	–
Doer's % Adjectives	0.10	0.07	–
Doer's % Verbs	-0.08	0.48	–
Doer's % Nouns	-0.06	0.37	–
Doer's % Adverbs	-0.03	0.09	–
Doer's % Questions	-0.07	0.55	–
Doer's % Continuers	-0.01	0.98	–

Table B.67: *Correlations - Doer's Values of Linguistic Measures as Percentages With Objective Conversation Quality - Aphasic Doer Cohort*

When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

B.5.11 Write-It Do-It: Correlations between Doer's Values of Linguistic Measures as Percentages and Writer's ICR

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Doer's % Function Words	<-0.01	0.48	–
Doer's % Content Words	<0.01	0.48	–
Doer's % Adjectives	0.02	0.01	0.41
Doer's % Verbs	<0.01	0.52	–
Doer's % Nouns	-0.01	0.05	-0.33
Doer's % Adverbs	<-0.01	0.97	–
Doer's % Questions	-0.02	0.17	–
Doer's % Continuers	-0.07	0.01	-0.40

Table B.68: *Correlations - Doer's Values of Linguistic Measures as Percentages With Writer's ICR Conversation Quality - Control Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Doer's % Function Words	-0.01	0.05	-0.32
Doer's % Content Words	0.01	0.05	0.32
Doer's % Adjectives	<-0.01	0.84	–
Doer's % Verbs	0.01	0.18	–
Doer's % Nouns	0.01	0.29	–
Doer's % Adverbs	<-0.01	0.46	–
Doer's % Questions	-0.02	0.44	–
Doer's % Continuers	-0.01	0.67	–

Table B.69: *Correlations - Doer's Values of Linguistic Measures as Percentages With Writer's ICR Conversation Quality - Aphasic Doer Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	**p-value**	**R^2**
Doer's % Function Words	0.01	0.22	–
Doer's % Content Words	-0.01	0.22	–
Doer's % Adjectives	0.01	0.53	–
Doer's % Verbs	-0.03	0.14	–
Doer's % Nouns	0.01	0.43	–
Doer's % Adverbs	<-0.01	0.67	–
Doer's % Questions	<-0.01	0.99	–
Doer's % Continuers	-0.2	0.70	–

Table B.70: *Correlations - Doer's Values of Linguistic Measures as Percentages With Writer's ICR Conversation Quality - Aphasic Doer Cohort*

When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

B.5.12 Write-It Do-It: Correlations between Doer's Values of Linguistic Measures as Percentages and Doer's ICR

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Doer's % Function Words	-0.01	0.34	–
Doer's % Content Words	0.01	0.34	–
Doer's % Adjectives	0.02	< 0.01	0.51
Doer's % Verbs	-0.01	< 0.01	-0.46
Doer's % Nouns	0.01	0.41	–
Doer's % Adverbs	<-0.01	0.28	–
Doer's % Questions	0.03	0.06	–
Doer's % Continuers	-0.09	< 0.01	-0.52

Table B.71: *Correlations - Doer's Values of Linguistic Measures as Percentages With Doer's ICR Conversation Quality - Control Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Doer's % Function Words	-0.01	0.15	–
Doer's % Content Words	0.01	0.15	–
Doer's % Adjectives	<-0.01	0.39	–
Doer's % Verbs	<0.01	0.52	–
Doer's % Nouns	0.01	0.26	–
Doer's % Adverbs	<0.01	0.45	–
Doer's % Questions	-0.05	0.10	–
Doer's % Continuers	0.02	0.33	–

Table B.72: *Correlations - Doer's Values of Linguistic Measures as Percentages With Doer's ICR Conversation Quality - Aphasic Doer Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	**p-value**	**R^2**
Doer's % Function Words	0.01	0.35	–
Doer's % Content Words	-0.01	0.35	–
Doer's % Adjectives	-0.01	0.47	–
Doer's % Verbs	-0.01	0.74	–
Doer's % Nouns	<0.01	0.76	–
Doer's % Adverbs	<0.01	0.90	–
Doer's % Questions	0.02	0.22	–
Doer's % Continuers	0.01	0.75	–

Table B.73: *Correlations - Doer's Values of Linguistic Measures as Percentages With Doer's ICR Conversation Quality - Aphasic Doer Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

B.5.13 Write-It Do-It: Correlations between Doer's Values of Linguistic Measures as Percentages and Writer's ICSI

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Doer's % Function Words	0.01	0.04	0.34
Doer's % Content Words	-0.01	0.04	-0.34
Doer's % Adjectives	-0.01	< 0.01	-0.51
Doer's % Verbs	<0.01	0.91	–
Doer's % Nouns	<0.01	0.55	–
Doer's % Adverbs	<0.01	0.96	–
Doer's % Questions	-0.01	0.27	–
Doer's % Continuers	0.05	0.04	0.34

Table B.74: *Correlations - Doer's Values of Linguistic Measures as Percentages With Writer's ICSI Conversation Quality - Control Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Doer's % Function Words	0.01	0.34	–
Doer's % Content Words	-0.01	0.34	–
Doer's % Adjectives	<0.01	0.59	–
Doer's % Verbs	-0.01	0.26	–
Doer's % Nouns	-0.01	0.51	–
Doer's % Adverbs	<0.01	0.40	–
Doer's % Questions	0.01	0.69	–
Doer's % Continuers	0.03	0.10	–

Table B.75: *Correlations - Doer's Values of Linguistic Measures as Percentages With Writer's ICSI Conversation Quality - Aphasic Doer Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Doer's % Function Words	<0.01	0.64	–
Doer's % Content Words	<-0.01	0.1	–
Doer's % Adjectives	-0.02	0.01	-0.42
Doer's % Verbs	0.03	0.10	–
Doer's % Nouns	<-0.01	0.79	–
Doer's % Adverbs	<-0.01	0.05	-0.33
Doer's % Questions	<0.01	0.95	–
Doer's % Continuers	0.04	0.26	–

Table B.76: *Correlations - Doer's Values of Linguistic Measures as Percentages With Writer's ICSI Conversation Quality - Aphasic Doer Cohort*

When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

B.5.14 Write-It Do-It: Correlations between Doer's Values of Linguistic Measures as Percentages and Doer's ICSI

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Doer's % Function Words	0.01	0.15	–
Doer's % Content Words	-0.01	0.15	–
Doer's % Adjectives	-0.01	0.01	-0.44
Doer's % Verbs	0.01	0.15	–
Doer's % Nouns	<-0.01	0.69	–
Doer's % Adverbs	<0.01	0.68	–
Doer's % Questions	-0.02	0.09	–
Doer's % Continuers	0.07	0.01	0.42

Table B.77: *Correlations - Doer's Values of Linguistic Measures as Percentages With Doer's ICSI Conversation Quality - Control Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	p-value	R^2
Doer's % Function Words	0.01	0.21	–
Doer's % Content Words	-0.01	0.21	–
Doer's % Adjectives	0.01	0.19	–
Doer's % Verbs	-0.01	0.08	–
Doer's % Nouns	<-0.01	0.74	–
Doer's % Adverbs	<-0.01	0.06	–
Doer's % Questions	0.03	0.30	–
Doer's % Continuers	-0.02	0.37	–

Table B.78: *Correlations - Doer's Values of Linguistic Measures as Percentages With Doer's ICSI Conversation Quality - Aphasic Doer Cohort*
When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

	GEE		Pearson's Correlation
	Coefficient	**p-value**	**R^2**
Doer's % Function Words	-0.01	0.38	–
Doer's % Content Words	0.01	0.38	–
Doer's % Adjectives	<0.01	0.73	–
Doer's % Verbs	0.02	0.25	–
Doer's % Nouns	-0.01	0.36	–
Doer's % Adverbs	<-0.01	0.21	–
Doer's % Questions	-0.03	0.16	–
Doer's % Continuers	-0.03	0.35	–

Table B.79: *Correlations - Doer's Values of Linguistic Measures as Percentages With Doer's ICSI Conversation Quality - Aphasic Doer Cohort*

When significance is found, we present the two R^2 values of the secondary analysis using Pearsons's Correlation, otherwise – is present.

B.6 Write-It Do-It: Order Effect Tables

B.6.1 Write-It Do-It: Summary Statistics – by First Conversations

	Mean (sd)		Median [IQ Range]		
Obj Score	0.32	(0.14)	0.32	[0.22 , 0.43]	
Writer ICSI	5.09	(1.27)	5.33	[4.67 , 5.92]	
Writer ICR	3.99	(1.10)	3.75	[3.30 , 4.30]	
Doer ICSI	5.25	(1.41)	5.42	[4.17 , 6.42]	
Writer ICR	3.95	(1.36)	3.70	[2.90 , 5.25]	

Table B.80: *Summary Statistics – First Conversation – Conversation Quality – Control Cohort*

	Mean (sd)		Median [IQ Range]		
Writer # Lines	81.56	(39.93)	76.00	[56.00 , 110.50]	
Writer # Words	760.69	(181.47)	724.00	[660.0 , 842.0( ]	
Writer % Function Words	0.43	(0.02)	0.43	[0.42 , 0.44]	
Writer % Content Words	0.57	(0.02)	0.57	[0.56 , 0.58]	
Writer % Adjectives	0.11	(0.02)	0.11	[0.10 , 0.12]	
Writer % Verbs	0.14	(0.02)	0.13	[0.12 , 0.15]	
Writer % Nouns	0.28	(0.03)	0.27	[0.26 , 0.28]	
Writer % Adverbs	0.05	(0.01)	0.05	[0.04 , 0.06]	
Writer % Questions	0.08	(0.06)	0.06	[0.04 , 0.12]	
Writer % Continuers	0.06	(0.06)	0.05	[0.02 , 0.08]	
Doer # Lines	50.25	(18.31)	47.00	[34.50 , 66.00]	
Doer # Words	303.13	(103.6( )	297.50	[229.0(, 383.0( ]	
Doer % Function Words	0.40	(0.04)	0.40	[0.38 , 0.43]	
Doer % Content Words	0.60	(0.04)	0.60	[0.57 , 0.62]	
Doer % Adjectives	0.11	(0.03)	0.11	[0.08 , 0.13]	
Doer % Verbs	0.17	(0.03)	0.17	[0.14 , 0.19]	
Doer % Nouns	0.27	(0.03)	0.27	[0.24 , 0.30]	
Doer % Adverbs	0.05	(0.02)	0.05	[0.04 , 0.06]	
Doer % Questions	0.31	(0.10)	0.31	[0.22 , 0.40]	
Doer % Continuers	0.29	(0.19)	0.29	[0.16 , 0.44]	

Table B.81: *Summary Statistics – First Conversation – Linguistic Measures – Control Cohort*
All values are % (in decimal form) of all Words that are of a specific type
† is percentage (in decimal form) of type per line

	Mean (sd)		Median [IQ Range]	
Obj Score	0.17	(0.06)	0.17	[0.15 , 0.22]
Writer ICSI	4.39	(1.29)	4.42	[3.33 , 5.33]
Writer ICR	4.95	(1.07)	4.80	[4.15 , 5.90]
Doer ICSI	4.58	(1.30)	4.58	[3.83 , 5.50]
Doer ICR	4.81	(1.30)	4.80	[4.05 , 5.45]

Table B.82: *Summary Statistics – First Conversation – Conversation Quality – Aphasic Writer Cohort*

	Mean (sd)		Median [IQ Range]	
Writer # Lines	82.13	(45.89)	70.00	[46.00 , 105.0( ]
Writer # Words	634.00	(268.3( )	565.00	[395.50 , 859.0( ]
Writer % Function Words	0.43	(0.03)	0.44	[0.41 , 0.45]
Writer % Content Words	0.57	(0.03)	0.56	[0.55 , 0.59]
Writer % Adjectives	0.11	(0.03)	0.11	[0.09 , 0.13]
Writer % Verbs	0.12	(0.02)	0.13	[0.10 , 0.13]
Writer % Nouns	0.30	(0.03)	0.30	[0.28 , 0.32]
Writer % Adverbs	0.04	(0.01)	0.04	[0.29 , 0.04]
Writer % Questions	0.07	(0.07)	0.05	[0.02 , 0.09]
Writer % Continuers	0.04	(0.05)	0.03	[0.01 , 0.05]
Doer # Lines	57.13	(44.50)	39.50	[31.50 , 67.00]
Doer # Words	336.56	(172.12)	348.00	[185.0(, 416.5( ]
Doer % Function Words	0.40	(0.06)	0.41	[0.39 , 0.42]
Doer % Content Words	0.60	(0.06)	0.59	[0.58 , 0.61]
Doer % Adjectives	0.10	(0.03)	0.10	[0.08 , 0.12]
Doer % Verbs	0.17	(0.03)	0.17	[0.15 , 0.19]
Doer % Nouns	0.29	(0.06)	0.28	[0.25 , 0.32]
Doer % Adverbs	0.04	(0.01)	0.04	[0.03 , 0.05]
Doer % Questions	0.48	(0.20)	0.44	[0.34 , 0.68]
Doer % Continuers	0.22	(0.11)	0.27	[0.17 , 0.29]

Table B.83: *Summary Statistics — First Conversation — Linguistic Measures — Aphasic Writer Cohort*
All values are % (in decimal form) of all Words that are of a specific type
† is percentage (in decimal form) of type per line

	Mean (sd)		Median [IQ Range]	
Obj Score	0.31	(0.18)	0.32	[0.15 , 0.39]
Writer ICSI	4.73	(1.10)	4.92	[3.75 , 5.58]
Writer ICR	4.39	(1.18)	4.40	[3.55 , 4.95]
Doer ICSI	4.74	(1.34)	4.92	[3.83 , 5.75]
Doer ICR	4.47	(0.89)	4.50	[3.75 , 5.10]

Table B.84: *Summary Statistics — First Conversation — Conversation Quality — Aphasic Doer Cohort*

	Mean (sd)		Median [IQ Range]	
Writer # Lines	77.38	(29.10)	78.00	[52.00 , 95.50]
Writer # Words	703.63	(242.5( )	692.50	[518.50 , 904.5( ]
Writer % Function Words	0.42	(0.03)	0.43	[0.41 , 0.45]
Writer % Content Words	0.58	(0.03)	0.57	[0.55 , 0.59]
Writer % Adjectives	0.11	(0.03)	0.11	[0.10 , 0.12]
Writer % Verbs	0.14	(0.02)	0.14	[0.12 , 0.15]
Writer % Nouns	0.28	(0.03)	0.28	[0.27 , 0.29]
Writer % Adverbs	0.05	(0.01)	0.04	[0.03 , 0.05]
Writer % Questions	0.10	(0.06)	0.07	[0.06 , 0.14]
Writer % Continuers	0.10	(0.08)	0.07	[0.04 , 0.17]
Doer # Lines	61.06	(25.93)	57.5	[38.50 , 76.50]
Doer # Words	284.56	(124.8( )	247.00	[194.50 , 391.0( ]
Doer % Function Words	0.40	(0.04)	0.41	[0.38 , 0.42]
Doer % Content Words	0.60	(0.04)	0.59	[0.58 , 0.62]
Doer % Adjectives	0.12	(0.06)	0.11	[0.07 , 0.17]
Doer % Verbs	0.17	(0.03)	0.18	[0.16 , 0.19]
Doer % Nouns	0.26	(0.04)	0.27	[0.22 , 0.28]
Doer % Adverbs	0.05	(0.02)	0.04	[0.03 , 0.06]
Doer % Questions	0.34	(0.11)	0.33	[0.27 , 0.42]
Doer % Continuers	0.16	(0.13)	0.12	[0.09 , 0.22]

Table B.85: *Summary Statistics — First Conversation — Linguistic Measures — Aphasic Doer Cohort*
All values are % (in decimal form) of all Words that are of a specific type
† is percentage (in decimal form) of type per line

B.6.2 Write-It Do-It: Summary Statistics – by Second Conversations

	Mean (sd)		Median [IQ Range]		
Obj Score	0.31	(0.15)	0.31	[0.23 , 0.44]	
Writer ICSI	5.20	(1.57)	5.42	[4.08 , 6.50]	
Writer ICR	3.59	(1.13)	3.65	[2.70 , 4.45]	
Doer ICSI	5.71	(1.05)	5.83	[5.42 , 6.33]	
Doer ICR	3.53	(0.93)	3.15	[2.80 , 4.30]	

Table B.86: *Summary Statistics - Second Conversation - Conversation Quality - Control Cohort .*

	Mean (sd)		Median [IQ Range]		
Writer # Lines	87.50	(42.41)	82.50	[55.00 , 122.50]	
Writer # Words	834.50	(163.0	851.00	[711.50 , 950.5	]
Writer % Function Words	0.43	(0.03)	0.43	[0.42 , 0.45]	
Writer % Content Words	0.57	(0.03)	0.57	[0.55 , 0.58]	
Writer % Adjectives	0.12	(0.03)	0.12	[0.09 , 0.14]	
Writer % Verbs	0.12	(0.02)	0.12	[0.11 , 0.14]	
Writer % Nouns	0.28	(0.02)	0.28	[0.27 , 0.30]	
Writer % Adverbs	0.05	(0.01)	0.04	[0.04 , 0.05]	
Writer % Questions	0.04	(0.04)	0.04	[0.02 , 0.05]	
Writer % Continuers	0.06	(0.06)	0.04	[0.02 , 0.08]	
Doer # Lines	65.19	(32.93)	61.00	[40.50 , 94.50]	
Doer # Words	262.63	(102.9	264.50	[198.0 , 342.5	]
Doer % Function Words	0.38	(0.37)	0.38	[0.35 , 0.41]	
Doer % Content Words	0.62	(0.04)	0.62	[0.59 , 0.65]	
Doer % Adjectives	0.10	(0.04)	0.10	[0.73 , 0.11]	
Doer % Verbs	0.19	(0.03)	0.19	[0.17 , 0.22]	
Doer % Nouns	0.28	(0.05)	0.27	[0.25 , 0.32]	
Doer % Adverbs	0.05	(0.02)	005	[0.04 , 0.06]	
Doer % Questions	0.31	(0.12)	0.28	[0.21 , 0.38]	
Doer % Continuers	0.29	(0.21)	0.31	[0.10 , 0.45]	

Table B.87: *Summary Statistics - Second Conversation - Linguistic Measures - Control Cohort*
All values are % (in decimal form) of all Words that are of a specific type
† is percentage (in decimal form) of type per line

	Mean (sd)			Median [IQ Range]			
Obj Score	0.17	(	0.06)	0.18	[0.13	, 0.22	]
Writer ICSI	4.68	(	1.16)	4.75	[3.83	, 5.58	]
Writer ICR	4.78	(	1.34)	4.65	[3.95	, 5.70	]
Doer ICSI	4.69	(	1.37)	4.67	[3.58	, 5.83	]
Doer ICR	4.78	(	1.02)	5.05	[4.00	, 5.35	]

Table B.88: *Summary Statistics – Second Conversation – Conversation Quality – Aphasic Writer Cohort*

	Mean (sd)			Median [IQ Range]			
Writer # Lines	116.00	(	59.08)	97.00	[76.50	, 148.50	]
Writer # Words	570.38	(	219.9( )	543.00	[433.50	, 723.0(	]
Writer % Function Words	0.37	(	0.07)	0.39	[0.31	, 0.43	]
Writer % Content Words	0.63	(	0.07)	0.61	[0.57	, 0.69	]
Writer % Adjectives	0.14	(	0.03)	0.14	[0.11	, 0.17	]
Writer % Verbs	0.11	(	0.02)	0.11	[0.09	, 0.13	]
Writer % Nouns	0.34	(	0.06)	0.32	[0.30	, 0.38	]
Writer % Adverbs	0.04	(	0.01)	0.04	[0.03	, 0.05	]
Writer % Questions	0.03	(	0.03)	0.02	[0.001	, 0.06	]
Writer % Continuers	0.03	(	0.03)	0.02	[0.00	, 0.05	]
Doer # Lines	79.81	(	37.22)	66.00	[54.50	, 95.50	]
Doer # Words	417.06	(	208.6)	377.50	[305.5(	, 557.50	]
Doer % Function Words	0.39	(	0.03)	0.39	[0.37	, 0.42	]
Doer % Content Words	0.61	(	0.03)	0.61	[0.58	, 0.63	]
Doer % Adjectives	0.12	(	0.04)	0.12	[0.10	, 0.15	]
Doer % Verbs	0.16	(	0.03)	0.15	[0.14	, 0.18	]
Doer % Nouns	0.29	(	0.05)	0.28	[0.26	, 0.31	]
Doer % Adverbs	0.04	(	0.01)	0.04	[0.02	, 0.05	]
Doer % Questions	0.50	(	0.18)	0.51	[0.42	, 0.59	]
Doer % Continuers	0.26	(	0.17)	0.28	[0.13	, 0.39	]

Table B.89: *Summary Statistics – Second Conversation – Linguistic Measures – Aphasic Writer Cohort*
All values are % (in decimal form) of all Words that are of a specific type
† is percentage (in decimal form) of type per line

	Mean (sd)			Median [IQ Range]			
Obj Score	0.31	(	0.18)	0.31	[0.15	, 0.39	]
Writer ICSI	5.26	(	1.01)	5.17	[4.58	, 6.25	]
Writer ICR	4.12	(	0.68)	4.20	[3.50	, 4.75	]
Doer ICSI	5.08	(	1.17)	5.33	[4.33	, 6.00	]
Doer ICR	3.78	(	1.28)	3.68	[3.35	, 4.25	]

Table B.90: *Summary Statistics – Second Conversation – Conversation Quality – Aphasic Doer Cohort*

	Mean (sd)		Median [IQ Range]	
Writer # Lines	85.88	(50.71)	76.00	[57.50 , 99.50]
Writer # Words	840.13	(245.3{)	769.50	[645.0(, 1006.( ]
Writer % Function Words	0.43	(0.02)	0.44	[0.43 , 0.45]
Writer % Content Words	0.56	(0.02)	0.56	[0.55 , 0.57]
Writer % Adjectives	0.12	(0.02)	0.11	[0.10 , 0.13]
Writer % Verbs	0.13	(0.02)	0.13	[0.11 , 0.15]
Writer % Nouns	0.27	(0.03)	0.27	[0.25 , 0.28]
Writer % Adverbs	0.05	(0.01)	0.06	[0.04 , 0.06]
Writer % Questions	0.07	(0.06)	0.06	[0.02 , 0.13]
Writer % Continuers	0.04	(0.04)	0.02	[0.01 , 0.06]
Doer # Lines	51.44	(18.72)	47.00	[38.50 , 70.50]
Doer # Words	171.75	(98.78)	160.00	[90.00 , 244.0( ]
Doer % Function Words	0.37	(0.07)	0.38	[0.37 , 0.40]
Doer % Content Words	0.63	(0.07)	0.62	[0.60 , 0.63]
Doer % Adjectives	0.09	(0.06)	0.08	[0.05 , 0.10]
Doer % Verbs	0.23	(0.15)	0.19	[0.16 , 0.24]
Doer % Nouns	0.28	(0.09)	0.28	[0.22 , 0.34]
Doer % Adverbs	0.04	(0.02)	0.04	[0.02 , 0.05]
Doer % Questions	0.21	(0.13)	0.21	[0.12 , 0.29]
Doer % Continuers	0.39	(0.22)	0.43	[0.21 , 0.55]

Table B.91: *Summary Statistics – Second Conversation – Linguistic Measures – Aphasic Doer Cohort*
All values are % (in decimal form) of all Words that are of a specific type
† is percentage (in decimal form) of type per line

B.6.3 Write-It Do-It: Comparative Statistics of Conversation Quality and Linguistic Measure Percentages between First and Second Conversation

	Control	Aphasic Writer	Aphasic Doer
Obj Score	0.95	0.97	0.97
Writer ICSI	0.83	0.49	0.14
Writer ICR	0.29	0.68	0.40
Doer ICSI	0.28	0.82	0.42
Doer ICR	0.29	0.93	0.07
Writer # Lines	0.67	0.06	0.55
Writer # Words	0.21	0.45	0.10
Writer % Function Words	0.80	< 0.01	0.19
Writer % Content Words	0.80	< 0.01	0.19
Writer % Adjectives	0.49	< 0.01	0.56
Writer % Verbs	0.06	0.18	0.47
Writer % Nouns	0.59	< 0.01	0.29
Writer % Adverbs	0.94	0.90	0.08
Writer % Questions	0.03	0.03	0.15
Writer % Continuers	0.79	0.30	< 0.01
Doer # Lines	0.10	0.11	0.21
Doer # Words	0.25	0.22	< 0.01
Doer % Function Words	0.15	0.79	0.15
Doer % Content Words	0.15	0.79	0.15
Doer % Adjectives	0.45	0.07	0.09
Doer % Verbs	0.04	0.26	0.12
Doer % Nouns	0.68	0.92	0.45
Doer % Adverbs	0.96	0.21	0.06
Doer % Questions	0.86	0.78	< 0.01
Doer % Continuers	0.95	0.81	< 0.01

Table B.92: *Comparative Statistics Between the First and Second Conversation – Within Each Cohort*
All Statistics are p-value from GEE analysis of

	Control	Aphasic Writer	Aphasic Doer
Obj Score	–	–	–
Writer ICSI	–	–	–
Writer ICR	–	–	–
Doer ICSI	–	–	–
Doer ICR	–	–	–
Writer # Lines	–	–	–
Writer # Words	–	–	–
Writer % Function Words	–	▼	–
Writer % Content Words	–	▲	–
Writer % Adjectives	–	▲	–
Writer % Verbs	–	–	–
Writer % Nouns	–	▲	–
Writer % Adverbs	–	–	–
Writer % Questions	▼	▼	–
Writer % Continuers	–	–	▼
Doer # Lines	–	–	–
Doer # Words	–	–	▼
Doer % Function Words	–	–	–
Doer % Content Words	–	–	–
Doer % Adjectives	–	–	–
Doer % Verbs	▲	–	–
Doer % Nouns	–	–	–
Doer % Adverbs	–	–	–
Doer % Questions	–	–	▼
Doer % Continuers	–	–	▲

Table B.93: *Comparative Statistics Between the First and Second Conversation – Change in Direction from First to Second Conversation*

Arrows indicate direction of change from first to second conversation, dashes indicate non-significant change.Blue is significance $p \leq 0.05$, Red $p \leq 0.01$

B.7 Write-It Do-It: Turn Taking

B.7.1 Write-It Do-It: Summary Statistics - Turn Taking

	Mean (sd)			Median [IQ Range]			
P_{W2W}	0.43	(	0.17)	0.41	[0.30	,	0.53]
P_{D2D}	0.21	(	0.14)	0.21	[0.10	,	0.27]

Table B.94: *Summary Statistics for Turn Taking Probabilities – Control Cohort*

	Mean (sd)			Median [IQ Range]			
P_{W2W}	0.43	(	0.17)	0.41	[0.29	,	0.60]
P_{D2D}	0.19	(	0.14)	0.15	[0.08	,	0.31]

Table B.95: *Summary Statistics for Turn Taking Probabilities – Aphasic Writer Cohort*

	Mean (sd)			Median [IQ Range]			
P_{W2W}	0.42	(	0.17)	0.43	[0.30	,	0.52]
P_{D2D}	0.20	(	0.12)	0.18	[0.11	,	0.26]

Table B.96: *Summary Statistics for Turn Taking Probabilities – Aphasic Doer Cohort*

B.7.2 Write-It Do-It: Comparative Statistics of Turn Taking Across Cohorts

	Control vs. Aphasic Writer	Control vs. Aphasic Doer	Aphasic Writer vs. Aphasic Doer
P_{W2W}	0.91	0.85	0.76
P_{D2D}	0.70	0.77	0.91

Table B.97: *Comparative Statistics Between the Three Cohorts – Turn Taking*
All Statistics are p-value from GEE analysis

B.7.3 Write-It Do-It: Correlations between Turn Taking and Conversation Quality

	GEE	
	Coefficient	**p-value**
P_{W2W} vs. Objective	-0.04	0.84
P_{D2D} vs. Objective	-0.26	0.12
P_{W2W} vs. Writer ICR	0.03	0.30
P_{D2D} vs. Writer ICR	-0.01	0.76
P_{W2W} vs. Writer ICSI	-0.03	0.11
P_{D2D} vs. Writer ICSI	0.01	0.47
P_{W2W} vs. Doer ICR	-0.01	0.66
P_{D2D} vs. Doer ICR	0.04	0.06
P_{W2W} vs. Doer ICSI	<0.01	0.93
P_{D2D} vs. Doer ICSI	-0.02	0.29

Table B.98: *Correlations - Turn Taking With Conversation Quality - Control Cohort*

	GEE	
	Coefficient	**p-value**
P_{W2W} vs. Objective	0.28	0.58
P_{D2D} vs. Objective	0.04	0.92
P_{W2W} vs. Writer ICR	0.02	0.34
P_{D2D} vs. Writer ICR	0.03	0.22
P_{W2W} vs. Writer ICSI	-0.06	0.01
P_{D2D} vs. Writer ICSI	<-0.01	0.89
P_{W2W} vs. Doer ICR	0.03	0.31
P_{D2D} vs. Doer ICR	0.03	0.15
P_{W2W} vs. Doer ICSI	-0.02	0.45
P_{D2D} vs. Doer ICSI	-0.01	0.59

Table B.99: *Correlations - Turn Taking With Conversation Quality - Aphasic Writer Cohort*

	GEE	
	Coefficient	**p-value**
P_{W2W} vs. Objective	0.09	0.60
P_{D2D} vs. Objective	-0.01	0.90
P_{W2W} vs. Writer ICR	0.01	0.83
P_{D2D} vs. Writer ICR	0.27	0.22
P_{W2W} vs. Writer ICSI	0.02	0.54
P_{D2D} vs. Writer ICSI	-0.03	0.15
P_{W2W} vs. Doer ICR	-0.02	0.35
P_{D2D} vs. Doer ICR	0.01	0.55
P_{W2W} vs. Doer ICSI	0.02	0.35
P_{D2D} vs. Doer ICSI	-0.02	0.25

Table B.100: *Correlations - Turn Taking With Conversation Quality - Aphasic Doer Cohort*

Appendix C

Copyright Permission for Published Research

After contacting ACM Permissions, I was informed that "As an ACM author, you should be aware that you do not need to request specific permission to reuse the papers you authored in your dissertation, which certainly qualifies as a new work of your own, with proper credit acknowledgment." This further conforms to the *Rights Retained by Authors and Original Copyright Holders*[1] as published by ACM. Complying with this, I provide the full citation (as required) both inline and at the beginning of chapters/sections that are derived from a previously published work.

[1] http://www.acm.org/publications/copyright_policy#Retained accessed on March 6, 2012

References

AbleNet (2008). Bigmack communicator - red.

Abowd, G. D., Gauger, M., and Lachenmann, A. (2003). The family video archive: an annotation and browsing environment for home movies. In *5th ACM SIGMM international workshop on Multimedia information retrieval*, Berkeley, California. ACM.

Ahn, H. and Picard, R. (2006). Affective cognitive learning and decision making: The role of emotions. In *The 18th European Meeting on Cybernetics and Systems Research (EMCSR 2006)*. Citeseer.

Al Mahmud, A., Gerits, R., and Martens, J. (2010). Xtag: designing an experience capturing and sharing tool for persons with aphasia. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, pages 325–334. ACM.

Al Mahmud, A. and Martens, J. (2011). Understanding email communication of persons with aphasia. In *Extended abstracts on Human factors in computing systems*, pages 1195–1200. ACM.

Alberto, P. and Troutman, A. (2005). *Applied behavior analysis for teachers*. Prentice Hall, Upper Saddle River, NJ.

Allen, J. and Core, M. (1997). Draft of damsl: Dialog act markup in several layers. *Unpublished manuscript*, 2.

Allen, M., McGrenere, J., and Purves, B. (2007). The design and field evaluation of phototalk: a digital image communication application for people. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pages 187–194. ACM.

Allen, M., McGrenere, J., and Purves, B. (2008). The field evaluation of a mobile digital image communication application designed for people with aphasia. *ACM Transactions on Accessible Computing (TACCESS)*, 1(1):5.

Altman, D. (1991). *Practical Statistics/or Medical Research*. Chapman and Hall, London.

Attainment Company (2008). Gotalk.

Autism Society of America, ASA (2007). Autism society of america.

Bach, F. R. and Jordan, M. I. (2006). Learning spectral clustering, with application to speech separation. *J. Mach. Learn. Res.*, 7:1963–2001.

Baer, D., Wolf, M., and Risley, T. (1968). Some current dimensions of applied behavior analysis. *Journal of applied behavior analysis*, 1(1):91–97.

Baggs, A. (2007). In my language.

Bales, R., Cohen, S., and Williamson, S. (1979). *SYMLOG: A system for the multiple level observation of groups*, volume 67. Free Press New York.

Banerjee, S., Cohen, J., Quisel, T., Chan, A., Patodia, Y., Al-Bawab, Z., Zhang, R., Rybski, P., Veloso, M., and Black, A. (2004). Creating multi-modal, userâĂŞcentric records of meetings with the carnegie mellon meeting recorder architecture. In *ICASSP Meeting Recognition Workshop*, Montreal, Quebec, Canada.

Barlow, D. H. and Hayes, S. C. (1979). Alternating treatments design: one strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis*, 12(2):199–210.

Barry, R. M. (1989). Epg from square one: An overview of electropalatography as an aid to therapy. *Clinical Linguistics & Phonetics*, 3(1):81–91.

Baskett, C. B. (1996). *The effect of live interactive video on the communicative behavior in children with autism*. PhD thesis, UNC - Chapel Hill.

Benjamin, C., Harris, J., Moncrief, A., Ramsberger, G., and Lewis, C. (2008). Naming practice on an open platform for people with aphasia. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*, pages 265–266. ACM.

Benson, D. (1979). *Aphasia, Alexia and Agraphia: Clinical Neurology and Neurosurgery Monographs*. Churchill Livingstone, New York.

Benson, D. (1994). *The neurology of thinking*. Oxford University Press, USA.

Bergstrom, T. and Karahalios, K. (2007a). Conversation clock: Visualizing audio patterns in co-located groups. In *HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCE 2007*, volume 40, Hawaii, USA. IEEE.

Bergstrom, T. and Karahalios, K. (2007b). Conversation votes: enabling anonymous cues. In *CHI '07 extended abstracts on Human factors in computing systems*, San Jose, CA, USA. ACM.

Bergstrom, T. and Karahalios, K. (2007c). Seeing more: Visualizing audio cues. In *INTERACT*, Rio de Janerio, Brasil. ACM Press.

Berry, K. and Mielke Jr, P. (1988). A generalization of cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement*, 48(4):921.

Bertini, M., Del Bimbo, A., Cucchiara, R., and Prati, A. (2004). Semantic video adaptation based on automatic annotation of sport videos. In *6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 291–298, New York, NY. ACM New York, NY, USA.

Beun, R. and Cremers, A. (1998). Object reference in a shared domain of conversation. *Pragmatics &# 38; Cognition, 6*, 1(2):121–152.

Bijou, S., Peterson, R., and Ault, M. (1968). A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. *Journal of Applied Behavior Analysis*, 1(2):175.

Birkimer, J. C. and Brown, J. H. (1979). A graphical judgmental aid which summarizes obtained and chance reliability data and helps assess the believability of experimental effects. *Journal of Applied Behavior Analysis*, 12(4):523–533.

Boden, M. (2006). *Mind as machine: a history of cognitive science*, volume 1. Clarendon press.

Boller, F. and Grafman, J. (2001). *Handbook of Neuropsychology, 2nd Edition : Language and Aphasia*. Elsevier Science Health Science div.

Bondy, A. and Frost, L. (2001). The picture exchange communication system. *Behavior Modification*, 25(5):725–744.

Bonneh, Y. S., Belmonte, M. K., Pei, F., Iversen, P. E., Kenet, T., Akshoomoff, N., Adini, Y., Simon, H. J., Moore, C. I., Houde, J. F., and Merzenich, M. M. (2008). Cross-modal extinction in a boy with severely autistic behaviour and high verbal intelligence. *Cognitive Neuropsychology*, 25(5):635 – 652.

Bos, N., Olson, J., Gergle, D., Olson, G., and Wright, Z. (2002). Effects of four computer-mediated communications channels on trust development. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*, pages 135–140. ACM.

Boyd-Graber, J., Nikolova, S., Moffatt, K., Kin, K., Lee, J., Mackey, L., Tremaine, M., and Klawe, M. (2006a). Participatory design with proxies: developing a desktop-pda system to support people with aphasia. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 151–160. ACM.

Boyd-Graber, J., Nikolova, S., Moffatt, K., Kin, K., Lee, J., Mackey, L., Tremaine, M., and Klawe, M. (2006b). Participatory design with proxies: developing a desktop-pda system to support people with aphasia. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 151–160. ACM.

Bradford-Heit, A. and Dodd, B. (1998). Learning new words using imitation and additional cues: differences between children with disordered speech. *Child Language Teaching & Therapy*, 14(2):159.

Brett, T. (2009). The challenges of game and human computer interface design for the rehabilitation of aphasic patients. In *Proceedings of the 2009 Conference on Future Play on GDC Canada*, pages 21–22. ACM.

Brown-Schmidt, S. and Tanenhaus, M. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive science*, 32(4):643–684.

Bryson, S. (1996). Brief report: Epidemiology of autism. *Journal of Autism and Developmental Disorders*, 26(2):165–167.

Burr, B. (2006). Vaca: a tool for qualitative video analysis. In *CHI '06 extended abstracts on Human factors in computing systems*, Montreal, Quebec, Canada. ACM.

Byrt, T. (1996). How good is that agreement? *Epidemiology*, 7(5):561.

Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J., and Anderson, A. (1997). The reliability of a dialogue structure coding scheme. *Computational linguistics*, 23(1):13–31.

Carter, P. and Edwards, S. (2004). Epg therapy for children with long-standing speech disorders: predictions and outcomes. *Clinical Linguistics & Phonetics*, 18(6):359.

Cassell, J., Kopp, S., Tepper, P., Ferriman, K., and Striegnitz, K. (2007). Trading spaces: How humans and humanoids use speech and gesture to give directions. In Nishida, I. T., editor, *Conversational Informatics*, pages 133–160. John Wiley & Sons, New York.

Center for Disease Control and Prevention, CDC (April 25, 2007). Autism information center.

Chandler, S., Harris, J., Moncrief, A., and Lewis, C. (2009). Naming practice for people with aphasia as a mobile web application. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, pages 247–248. ACM.

Chang, E. and Ma, C. (2002). Implicit speech recognition: making speech a first class object on computers. In *Proceedings of the second international conference on Human Language Technology Research*, San Diego, California. Morgan Kaufmann Publishers Inc.

Chao, D. (2001). Doom as an interface for process management. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 152–157. ACM.

Chapey, R. (2001). *Language intervention strategies in aphasia and related neurogenic communication disorders*. Lippincott Williams & Wilkins.

Chen, S., Shyu, M., Liao, W., and Zhang, C. (2002). Scene change detection by audio and video clues. In *Multimedia and Expo, 2002. ICME'02*, volume 2, pages 365–368. IEEE.

Clark, H. and Schaefer, E. (1989). Contributing to discourse. *Cognitive science*, 13(2):259–294.

Clifford, J., Marcus, G. E., and of American Research, S. (1986). *Writing Culture*. University of California Press, Berkeley, California.

Clifton, R. K., Perrisa, E. E., and McCalla, D. D. (1999). Does reaching in the dark for unseen objects reflect representation in infants? *Infant Behavior and Development*, 22(3):297–302.

Colby, K., Hilf, F., Weber, S., and Kraemer, H. (1972). Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. *Artificial Intelligence*, 3:199–221.

Colby, K., Weber, S., and Hilf, F. (1971). Artificial paranoia. *Artificial Intelligence*, 2(1):1–25.

Coletto, M. (2009). Comparing forms of visual feedback to facilitate early word use in two preschool-age children. Master's thesis, University of Illinois at Urbana Champaign.

Conger, A. (1985). Kappa reliabilities for continuous behaviors and events. *Educational and Psychological Measurement*, 45(4):861.

Consolvo, S., Everitt, K., Smith, I., and Landay, J. (2006). Design requirements for technologies that encourage physical activity. In *Proceedings of the SIGCHI*, pages 457–466. ACM.

Core, M., Ishizaki, M., Moore, J., Nakatani, C., Reithinger, N., Traum, D., and Tutiya, S. (1998). The report of the third workshop of the discourse resource initiative, chiba university and kazusa academia hall.

Costa, M., Correia, N., Guimar, N., and es (2002). Annotations as multiple perspectives of video content. In *ACM international conference on Multimedia*, Juan-les-Pins, France. ACM.

Crane, E. A., Shami, N. S., and Peter, C. (2007). Let's get emotional: emotion research in human computer interaction. In *CHI '07 extended abstracts on Human factors in computing systems*, San Jose, CA, USA. ACM.

Cushman, W. and Rosenberg, D. (1991). Human factors in product design. *Advances in human factors/ergonomics*, 14.

Czvik, P. (1977). Assessment of family attitudes toward aphasic patients with severe auditory processing disorders. *Clinical Aphasiology Conference*.

Daemen, E., Dadlani, P., Du, J., Li, Y., Erik-Paker, P., Martens, J., and De Ruyter, B. (2010a). Designing a free style, indirect, and interactive storytelling application for people with aphasia. *Human-Computer Interaction–INTERACT 2007*, pages 221–234.

Daemen, E., Dadlani, P., Du, J., Li, Y., Erik-Paker, P., Martens, J., and De Ruyter, B. (2010b). Designing a free style, indirect, and interactive storytelling application for people with aphasia. *INTERACT 2007*, pages 221–234.

Danilevsky, M., Hailpern, J., and Han, J. (2011). SCENE: Structural Conversation Evolution NEtwork. In *Proceedings of Social Networks Analysis and Mining (ASONAM 2011)*. IEEE.

Davies, R., Marcella, S., McGrenere, J., and Purves, B. (2004). *The ethnographically informed participatory design of a PD application to support communication*. Number 77-78. ACM.

Dettmer, S., Simpson, R., Myles, B., and Ganz, J. (2000). The use of visual supports to facilitate transitions of students with autism. *Focus on Autism and Other Developmental Disabilities*, 15(3):163.

Devlin, S. and Unthank, G. (2006). Helping aphasic people process online information. In *Proceedings of the 8th international ACM SIGACCESS Conference on Computers and Accessibility*, pages 225–226. ACM.

Doerksen, S. (2009). Recreation for persons with disabilities (rpm 277). *Pennsylvania State University Department of Recreation, Park and Tourism Management*, Fall.

Dollaghan, C. and Campbell, T. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, 41(5):1136.

Duck, S., Rutt, D., HOY, M., and STREJC, H. (1991). Some evident truths about conversations in everyday relationships all communications are not created equal. *Human communication research*, 18(2):228–267.

Dymond, R. (1949). A scale for the measurement of empathic ability. *Journal of Consulting Psychology*, 13(2):127–133.

Edwards, W., Bellotti, V., Dey, A., and Newman, M. (2003). The challenges of user-centered design and evaluation for infrastructure. In *Proceedings of the SIGCHI*, pages 297–304. ACM.

El Kaliouby, R. and Robinson, P. (2004). Real-time inference of complex mental states from facial expressions and head gestures. In *Computer Society Conference on. Computer Vision and Pattern Recognition*, volume 3, page 154, Washington, DC. Springer.

Fell, H., MacAuslan, J., Gong, J., Cress, C., and Salvo, T. (2006). visibabble for pre-speech feedback. In *Extended abstracts of CHI'06*, pages 767–772. ACM.

Fenson, L., Dale, P., Reznick, J., Thal, D., Bates, E., Hartung, J., Pethick, S., Reilly, J., et al. (2002). *MacArthur Communicative Development Inventories: User's guide and technical manual*. Paul H. Brookes Baltimore, MD.

Field, T., Field, T., Sanders, C., and Nadel, J. (2001). Children with autism display more social behaviors after repeated imitation sessions. *Autism*, 5(3):317–323.

Fry, B. and Reas, C. (2007). Processing.

Furbacher, E. A. and Wertz, R. T. (1983). Simulation of aphasia by wives of aphasic patients. *Clinical Aphasiology*, page 227.

Gallagher, A., Laxon, V., Armstrong, E., and Frith, U. (1996). Phonological difficulties in high-functioning dyslexics. *Reading and Writing*, 8(6):499–509.

Gamma, E. (1995). *Design patterns: elements of reusable object-oriented software*. Addison-Wesley Professional.

Gathercole, S. and Baddeley, A. (1993). *Working Memory and Language*. Psychology Pr.

Gena, A., Krantz, P., McClannahan, L., and Poulson, C. (1996). Training and generalization of affective behavior displayed by youth with autism. *Journal of Applied Behavior Analysis*, 29(3):291–304.

General Magic (1994). Magiccap.

Giles, H. (1971). *A study of speech patterns in social interaction: Accent evaluation and accent change*. PhD thesis, University of Bristol, Bristol, UK.

Giles, H., Fortman, J., Dailey, R., Barker, V., Hajek, C., Anderson, M., and Rule, N. (2006). Communication accommodation: Law enforcement and the public. *Applied interpersonal communication matters: Family, health, and community relations*, pages 241–269.

Giles, H., Taylor, D. M., and Bourhis, R. (1973). Towards a theory of interpersonal accommodation through language: Some canadian data. *Language in Society*, 2(2):pp. 177–192.

Gillette, D., Hayes, G., Abowd, G., Cassell, J., el Kaliouby, R., Strickland, D., and Weiss, P. (2007). Interactive technologies for autism. In *CHI 07*, San Jose, CA. ACM.

Goodglass, H., Goodglass, and Kaplan (2001). *Boston Diagnostic Aphasia Examination: Stimulus Cards–Short Form*. Lippincott Williams & Wilkins.

Grandin, T. (2006). *Thinking in Pictures: And Other Reports from My Life with Autism*. Vintage Books, New York.

Greenspan, S. I. and Wieder, S. (1997). Developmental patterns and outcomes in infants and children with disorders in relating and communicating: A chart review of 200 cases of children with autistic spectrum diagnoses. *The Journal of Developmental and Learning Disorders*, 1(1).

Hagedorn, J., Hailpern, J., and Karahalios, K. G. (2008). Vcode and vdata: Illustrating a new framework for supporting the video annotation workïňĆow. In *Advanced Visual Interfaces*, Napoli, Italy. ACM-PRESS.

Hailpern, J. (2008). The Spoken Impact Project: Using Audio & Visual Feedback to Impact Vocalization in Non-Verbal Children with Autistic Spectrum Disorder. Master's thesis, Department of Computer Science at the University of Illinois at Urbana-Champaign, Urbana, IL.

Hailpern, J., Danilevsky, M., Harris, A., Karahalios, K., Dell, G., and Hengst, J. (2011a). Aces: promoting empathy towards aphasia through language distortion emulation software. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, New York, NY, USA. ACM.

Hailpern, J., Danilevsky, M., and Karahalios, K. (2011b). Aces: aphasia emulation, realism, and the turing test. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, New York, NY, USA. ACM.

Hailpern, J., Karahalios, K., DeThorne, L., and Halle, J. (2009a). Talking points: The differential impact of real-time computer generated audio/visual feedback on speech-like & non-speech-like vocalizations in low functioning children with asd. In *Proceedings of the ASSETS'09*, pages 187–194. ACM.

Hailpern, J., Karahalios, K., and Halle, J. (2009b). Creating a spoken impact: encouraging vocalization through audio visual feedback in children with asd. In *CHI 2009*, Boston, MA. ACM.

Hailpern, J., Karahalios, K., Halle, J., DeThorne, L., and Coletto, M. (2008). A3: A coding guideline for hci+autism research using video annotation. In *ACM SIGACCESS- ASSETS 2008*, Halifax, Canada. ACM-PRESS.

Hailpern, J., Karahalios, K., Halle, J., Dethorne, L., and Coletto, M.-K. (2009c). A3: Hci coding guideline for research using video annotation to assess behavior of nonverbal subjects with computer-based intervention. *ACM Trans. Access. Comput.*, 2.

Halle, J. (1987). Teaching language in the natural environment: An analysis of spontaneity. *Journal of the Association for Persons with Severe Handicaps (JASH)*, 12(1):28–37.

Hardin, J. and Hilbe, J. (2003). *Generalized estimating equations*. Chapman and Hall/CRC, New York.

Hartsuiker, R. and Kolk, H. (1998). Syntactic facilitation in agrammatic sentence production. *Brain and Language*, 62(2):221–254.

Hayden, D. (1999). *VMPAC: Verbal motor production assessment for children*. Psychological Corporation.

Hayes, G. R., Kientz, J. A., Truong, K. N., White, D. R., Abowd, G. D., and Pering, T. (2004). Designing capture applications to support the education of children with autism. In *International Conference on Ubiquitous Computing*, Nottingham, England.

Hayne, H., Gross, J., Hildreth, K., and Rovee-Collier, C. (2000). Repeated reminders increase the speed of memory retrieval by 3-month-old infants. *Developmental Science*, 3(3):312–318.

Hecht, M. (1978). The conceptualization and measurement of interpersonal communication satisfaction. *Human Communication Research*, 4(3):253–264.

Heer, J., Viegas, F. B., and Wattenberg, M. (2007). Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *Proceedings of SIGCHI*, San Jose, California, USA. ACM.

Hengst, J. (2005). Augmentative and alternative communication (shs 473). *University of Illinois Department of Speech and Hearing Science*, Spring.

Highman, C., Hennessey, N., Sherwood, M., and Leitão, S. (2008). Retrospective parent report of early vocal behaviours in children with suspected childhood apraxia of speech (scas). *Child Language Teaching and Therapy*, 24(3):285.

Hoque, M., Lane, J., El Kaliouby, R., Goodwin, M., and Picard, R. (2009). Exploring speech therapy games with children on the autism spectrum. In *Proceedings of InterSpeech*. Citeseer.

Howitt, A. W. (2000). *Automatic syllable detection for vowel landmarks*. Ph.d thesis. supervisor - kenneth n. stevens, Massachusetts Institute of Technology.

Howlin, P. (1986). An overview of social behavior in autism. In Schopler, E. and Mesibov, G. B., editors, *Social Behavior in Autism*, Current Issues in Autism, pages 113–115. Plenum, New York, NY.

IBM (1997). Speech Viewer III.

IBM (2005). Ibm viavoice.

Jacob, R. J. K. (1991). The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Trans. Inf. Syst.*, 9(2):152–169.

Jefferson, G. (1984). Notes on a systematic deployment of the acknowledgement tokens ÒyeahÓ; and ÒmmhmÓ.

Johnson, A. (2007). About vprism video data analysis software.

Jones, C. M. and Troen, T. (2007). Biometric valence and arousal recognition. In *Proceedings of the 2007 conference of the computer-human interaction special interest group (CHISIG) of Australia on Computer-human interaction: design: activities, artifacts and environments*, Adelaide, Australia. ACM.

Jones, V. and Prior, M. (1985). Motor imitation abilities and neurological signs in autistic children. *Journal of Autism and Developmental Disorders*, 15(1):37–46.

Jurafsky, D., Martin, J., Kehler, A., Vander Linden, K., and Ward, N. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, volume 163. MIT Press.

Kaliouby, R. e. and Teeters, A. (2007). Eliciting, capturing and tagging spontaneous facialaffect in autism spectrum disorder. In *Proceedings of the 9th international conference on Multimodal interfaces*, Nagoya, Aichi, Japan. ACM.

Kanner, L. (1943). Autistic disturbances of affective contact. In Kanner, L., editor, *Nervous Child 2*, pages 217–250. V.H. Winston.

Karahalios, K. and Donath, J. (2004). Telemurals: linking remote spaces with social catalysts. In *SIGCHI conference on Human factors in computing systems*, Vienna, Austria. ACM New York, NY, USA.

Karahalios, K. and Viegas, F. B. (1999). Visiphone. In *ACM SIGGRAPH 99 Conference abstracts and applications*, Los Angeles, California, United States. ACM.

Karahalios, K. G. and Dobson, K. (2005). Chit chat club: bridging virtual and physical space for social interaction. In *CHI '05 extended abstracts on Human factors in computing systems*, Portland, OR, USA. ACM.

Karahalios, K. G. and Viegas, F. B. (2006). Social visualization: exploring text, audio, and video interaction. In *CHI '06 extended abstracts on Human factors in computing systems*, Montrial, Quibec, Canada. ACM.

KayPentax (1996-2008). Visi-Pitch IV, Model 2950B.

Kazdin, A. (1977). Artifact, bias, and complexity of assessment: the abcs of reliability. *Journal of Applied Behavior Analysis*, 10(1):141.

Kazdin, A. E. (1982). *Single-Case Research Designs: Methods for Clinical and Applied Setting*. Oxford University Press, USA.

Kerr, S. J., Neale, H. R., and Cobb, S. V. G. (2002). Virtual environments for social skills training: the importance of scaffolding in practice. In *Proceedings of ASSETS'02*, Edinburgh, Scotland. ACM Press.

Kientz, J. A., Arriaga, R. I., Chetty, M., Hayes, G. R., Richardson, J., Patel, S. N., and Abowd, G. D. (2007). Grow and know: understanding record-keeping needs for tracking the development of young children. In *Proceedings of SIGCHI*, San Jose, California, USA. ACM Press.

Kientz, J. A., Hayes, G. R., Abowd, G. D., and Grinter, R. E. (2006). From the war room to the living room: decision support for home-based therapy teams. In *Proceedings of CSCW 2006*, Banff, Alberta, Canada. ACM Press.

Kipp (2007). Anvil - the video annotation research tool.

Kirvesoja, H. (2001). *Experimental ergonomic evaluation with user trials: EEE product development procedures*. Oulun yliopisto.

Kitzing, P., Maier, A., and Ąhlander, V. (2009). Automatic speech recognition (asr) and its use as a tool for assessment or therapy of voice, speech, and language disorders. *Logopedics Phonatrics Vocology*, 34(2):91–96.

Koegel, L. K., Koegel, R. L., Harrower, J. K., and Carter, C. M. (1999). Pivotal response intervention i: Overview of approach. *The Journal of The Association for Persons with Severe Handicaps*, 24(3):174–185.

Kolk, H. (1995). A time-based approach to agrammatic production. *Brain and Language*, 50(3):282–303.

Kolk, H. and Heeschen, C. (1990). Adaptation symptoms and impairment symptoms in broca's aphasia. *Aphasiology*, 4(3):221–231.

Kolk, H. and Heeschen, G. (1996). The malleability of agrammatic symptoms: A reply to hesketh and bishop. *Aphasiology*, 10(1):81–96.

Kolk, H. and Van Grunsven, M. (1985). Agrammatism as a variable phenomenon. *Cognitive Neuropsychology*.

Kwon, S. and Narayanan, S. (2004). Speaker model quantization for unsupervised speaker indexing. In *International Conference Spoken Language Processing*, Jeju, Korea.

Leadholm, B., Miller, J., and Children, B. f. E. (1992). *Language Sample Analysis: The Wisconsin Guide*. Wisconsin Dept. of Public Instruction.

Lee, L. (1974). *Developmental Sentence Analysis*. Northwestern University Press, Evanston, IL.

Lehman, J. F. (1998). Toward the use of speech and natural language technology in intervention for a language-disordered population. In *Proceedings of the ASSETS'98*, Marina del Rey, California, United States. ACM Press.

Leonard, L. B. (2000). *Children with Specific Language Impairment*. MIT Press, Cambridge, MA.

Leshed, G. (2009). *Automated language-based feedback for teamwork behaviors*. PhD thesis, Cornell University.

Leung, S. T. (2006). Integrating visualization to make programming concepts concrete: dot net style. In *Proceedings of the 7th conference on Information technology education*, Minneapolis, Minnesota, USA. ACM.

Levin, G. and Lieberman, Z. (2004). In-situ speech visualization in real-time interactive installation and performance. In *Non-Photorealistic Animation and Rendering*, volume 7, pages 7–14, Annecy, France. ACM.

Lewis, C. and Rieman, J. (1993). Task-centered user interface design. *Available via ftp. cs. colorado. edu/pub/cs/distribs/clewis/HCI-Design-Books*.

Liechty, J. and Heinzekehr, J. (2007). Caring for those without words: A perspective on aphasia. *The Journal of Neuroscience Nursing*, 39(5):316.

Liu, H., Aingh, P., and Eslick, I. (2010). Conceptnet.

Lord, C., Risi, S., Lambrecht, L., Cook, E. J., Leventhal, B., DiLavore, P., Pickles, A., and Rutter, M. (2000). The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, 30(3):205–223.

Lovaas, I. I. (1977). *The Autistic Child*. John Wiley & Sons, Inc, New York.

Lovaas, O. I. (2003). *Teaching Individuals with Developmental Delays: Basic Intervention Techniques.* PRO-ED, Inc., Austin, TX.

Luo, Y. and Baillargeon, R. (2005). Can a self-propelled box have a goal? *Psychological Science*, 16(8):601–608.

Ma, X., Boyd-Graber, J., Nikolova, S., and Cook, P. (2009). Speaking through pictures: Images vs. icons. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, pages 163–170. ACM.

MacWhinney, B. (1998). Models of the emergence of language. *Annual review of psychology*, 49(1):199–227.

Madsen, M., Kaliouby, R. e., Goodwin, M., and Picard, R. (2008). Technology for just-in-time in-situ learning of facial affect for persons diagnosed with an autism spectrum disorder. In *Proceedings of ASSETS'08*, Halifax, Nova Scotia, Canada. ACM.

Mahurin-Smith, J. (2010). *The impact of prematurity on language skills at school age*. PhD thesis, University of Illinois at Urbana Champaign.

Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S., and Ames, M. (2003). Heuristic evaluation of ambient displays. In *Proceedings of SIGCHI*, Ft. Lauderdale, Florida, USA. ACM.

Marshalla, P. (2001). *Becoming Verbal With Childhood Apraxia: New Insights on Piaget for Today's Therapy*. Marshalla Speech and Language, Kirkland, WA.

Martins, B. (2010). Jaspell.

Masson, V. and Quiniou, R. (1990). Application of artificial intelligence to aphasia treatment. In *Proceedings of the 3rd international conference on Industrial and engineering applications of artificial intelligence and expert systems-Volume 2*, pages 907–913. ACM.

McCall, D., Virata, T., Linebarger, M., and Berndt, R. (2009). Integrating technology and targeted treatment to improve narrative production in aphasia: A case study. *Aphasiology*, 23(4):438–461.

McCalla, D. D. and Clifton, R. K. (1999). Infants' means-end search for hidden objects in the absence of visual feedback. *Infant Behavior and Development*, 22(2):179–195.

McCann, R. and Giles, H. (2006). Communication with people of different ages in the workplace: Thai and American data. *Human communication research*, 32(1):74–108.

McGrenere, J., Davies, R., Findlater, L., Graf, P., Klawe, M., Moffatt, K., Purves, B., and Yang, S. (2003). Insights from the aphasia project: designing technology for and with people who have aphasia. In *ACM SIGCAPH Computers and the Physically Handicapped*, number 73-74, pages 112–118. ACM.

McQuiggan, S. and Lester, J. (2006). Learning empathy: A data-driven framework for modeling empathetic companion agents. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 961–968. ACM.

Mehrabian, A. and Epstein, N. (1972). A measure of emotional empathy. *Journal of personality*, 40(4):525–543.

Melder, W. A., Truong, K. P., Uyl, M. D., Leeuwen, D. A. V., Neerincx, M. A., Loos, L. R., and Plum, B. S. (2007). Affective multimodal mirror: sensing and eliciting laughter. In *Proceedings of the international workshop on Human-centered multimedia*, Augsburg, Bavaria, Germany. ACM.

Menn, L., Kamio, A., Hayashi, M., Fujita, I., Sasanuma, S., and Boles, L. (1998a). The role of empathy in sentence production: A functional analysis of aphasic and normal elicited narratives in Japanese and English. *Function and Structure*, pages 317–356.

Menn, L. and Obler, L. (1990). *Agrammatic aphasia: A cross-language narrative sourcebook*. John Benjamins.

Menn, L., Reilly, K., Hayashi, M., Kamio, A., Fujita, I., and Sasanuma, S. (1998b). The interaction of preserved pragmatics and impaired syntax in Japanese and English aphasic speech. *Brain and language*, 61(2):183–225.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *J. Acoust. Soc. Am*, 58(4):880–883.

Michaud, F. and Theberge-Turmel, C. (2002). Mobile robotic toys and autism. In Dautenhahn, K., editor, *Socially Intelligent Agents - Creating Relationships with Computers and Robots*, volume 3, pages 125–132. Springer.

Microsoft Corporation (1995). Bob.

Millar, D., Light, J., and Schlosser, R. (2006). The impact of augmentative and alternative communication intervention on the speech production of individuals with developmental disabilities: a research review. *Journal of Speech, Language, and Hearing Research*, 49(2):248.

Minshew, N. J., Goldstein, G., and Siegel, D. J. (1997). Neuropsychologic functioning in autism: Profile of a complex information processing disorder. *Journal of the International Neuropsychological Society*, 3.:303–316.

Moffatt, K., McGrenere, J., Purves, B., and Klawe, M. (2004a). The participatory design of a sound and image enhanced daily planner for people with aphasia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 407–414. ACM.

Moffatt, K., McGrenere, J., Purves, B., and Klawe, M. (2004b). The participatory design of a sound and image enhanced daily planner for people with aphasia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 407–414. ACM.

Mohamed, A. O., Courboulay, V., Sehaba, K., and Menard, M. (2006). Attention analysis in interactive software for children with autism. In *Proceedings of ASSETS'06*, Portland, Oregon. ACM.

Morris, R., Kirschbaum, C., and Picard, R. (2010). Broadening accessibility through special interests: a new approach for software customization. In *Proceedings of ASSETS'10*, pages 171–178. ACM.

Mukhopadhyay, T. R. (2000). *Beyond the Silence: My Life, the World and Autism*. National Autistic Society, London.

Mullennix, J. and Stern, S. (2010). *Computer Synthesized Speech Technologies: Tools for Aiding Impairment*. Medical Information Science Reference, Hershey, PA.

Nakagawa, S. (2002). A survey on automatic speech recognition. *IEICE TRANSACTIONS on Information and Systems*, E85-D(3):465–486.

National Institute on Deafness and Other Communication Disorders (2010). Aphasia. http://www.nidcd.nih.gov/health/voice/aphasia.htm.

National Research Council (2001). Educating children with autism. Technical report, Division of Behavioral and Social Sciences and Education.

Nguyen, D. and Rosé, C. (2011). Language use as a reflection of socialization in online communities. *ACL HLT 2011*, page 76.

Nikolova, S., Boyd-Graber, J., and Cook, P. (2009). The design of viva: a mixed-initiative visual vocabulary for aphasia. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pages 4015–4020. ACM.

Nikolova, S., Tremaine, M., and Cook, P. (2010). Click on bake to get cookies: guiding word-finding with semantic associations. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*, pages 155–162. ACM.

Nishida, M. and Ariki, Y. (1998). Real time speaker indexing based on subspace method-application to tv news articles and debate. In *Fifth International Conference on Spoken Language Processing*, Sydney, Australia.

Noldus (2007). Noldus. the observer.

Nuance Communications (2008). Dragon naturallyspeaking.

Olsson, R. K. and Hansen, L. K. (2006). Linear state-space models for blind source separation. *J. Mach. Learn. Res.*, 7:2585–2602.

Osadchy, M., Cun, Y. L., and Miller, M. L. (2007). Synergistic face detection and pose estimation with energy-based models. *J. Mach. Learn. Res.*, 8:1197–1215.

Owens, R. E. (2007). *Language Development: An Introduction (7th Edition)*. Allyn & Bacon, Boston, MA.

Pares, N., Carreras, A., Durany, J., Ferrer, J., Freixa, P., Gomez, D., Kruglanski, O., Pares, R., Ribas, J. I., Soler, M., and Sanjurjo, A. (2005). Promotion of creative activity in children with severe autism through visuals in an interactive multisensory environment. In *Proceeding of the 2005 conference on Interaction design and children*, Boulder, Colorado. ACM Press.

Paul, R. (2008). Recommendation for benchmarks - autism speaks luncheon at srcld.

Paul, R., Chawarska, K., Fowler, C., Cicchetti, D., and Volkmar, F. (2007). Listen my children and you shall hearâĂİ: Auditory preferences in toddlers with autism spectrum disorders. *Journal of Speech, Language, and Hearing Research*, 50.:1350–1364.

Pennebaker, J., Booth, R. J., and Francis, M. E. (2007). Linguistic inquiry and word count: Liwc.

Perlman, A. (2008). Speech and hearing science proseminar.

Peterson, S., Bondy, A., Vincent, Y., and Finnegan, C. (1995). Effects of altering communicative input for students with autism and no speech: Two case studies. *Augmentative and Alternative Communication*, 11(2):93–100.

Petridis, S. and Pantic, M. (2008). Fusion of audio and visual cues for laughter detection. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, Niagara Falls, Canada. ACM.

Pickering, M. and Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language & Computation*, 4(2):203–228.

Piper, A., O'Brien, E., Morris, M., and Winograd, T. (2006). Sides: a cooperative tabletop computer game for social skills development. In *CSCW 06*, Banff, Canada. ACM Press.

Piper, A., Weibel, N., and Hollan, J. (2010). Introducing multimodal paper-digital interfaces for speech-language therapy. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*, pages 203–210. ACM.

Plaue, C., Miller, T., and Stasko, J. (2004). Is a picture worth a thousand words?: an evaluation of information awareness displays. In *Proceedings of Graphics Interface 2004*, London, Ontario, Canada. Canadian Human-Computer Communications Society.

Ponceleon, D. and Srinivasan, S. (2001). Automatic discovery of salient segments in imperfect speech transcripts. In *tenth international conference on Information and knowledge management*, pages 490–497, Atlanta, Georgia. ACM New York, NY, USA.

Porter, M. F. (2001). Snowball: A language for stemming algorithms. Published online. Accessed 11.03.2008, 15.00h.

Preston, J. and Edwards, M. (2007). Phonological processing skills of adolescents with residual speech sound errors. *Language, speech, and hearing services in schools*, 38(4):297.

Prizant, B., Schuler, A., Wetherby, A. M., and Rydell, P. (1997). Enhancing language and communication: Language approaches. In Cohen, D. and Volkmar, F., editors, *Handbook of Autism and Pervasive Developmental Disorders (Second Edition)*. Wiley, New York.

Quek, F., McNeill, D., Rose, T., and Shi, Y. (2003). A coding tool for multimodal analysis of meeting video. In *NIST Meeting Room Workshop*, Montreal, Quebec, Canada.

Ramos, G. and Balakrishnan, R. (2003). Fluid interaction techniques for the control and annotation of digital video. In *Proceedings of the 16th annual ACM symposium on User interface software and technology*, Vancouver, Canada. ACM.

Rapin, I. and Dunn, M. (1997). Language disorders in children with autism. *Seminars in Pediatric Neurology*, 4(2):86–92.

Reichle, J., Beukelman, D., and Light, J. (2002). *Implementing an augmentative communication system: Exemplary strategies for beginning communicators*. Brookes Publishing Company, Baltimore, MD.

Reid, D. H., Parsons, M. B., McCarn, J. E., Green, C. W., Phillips, J. F., and Schepis, M. M. (1985). Providing a more appropriate education for severely handicapped persons: increasing and validating functional classroom tasks. *Journal of Applied Behavior Analysis*, 18(4):289–301.

Retherford, K., Sewards, D., and Hess, L. (1993). *Guide to Analysis of Language Transcripts*. Thinking Publications Eau Claire, WI.

Rita, H. and Ekholm, P. (2007). Showing similarity of results given by two methods: A commentary. *Environmental Pollution*, 145(2):383–386.

Roads, C., Strawn, J., Abbott, C., Gordon, J., and Greenspun, P. (1996). *The computer music tutorial*, volume 81. MIT press Cambridge, Massachusetts.

Rosenblum, K., Zeanah, C., McDonough, S., and Muzik, M. (2004). Video-taped coding of working model of the child interviews: a viable and useful alternative to verbatim transcripts? *Infant Behavior and Development*, 27(4):544–549.

Ruiter, M., Kolk, H., and Rietveld, T. (2010). Speaking in ellipses: The effect of a compensatory style of speech on functional communication in chronic agrammatism. *Neuropsychological rehabilitation*, 20(3):423–458.

Ruiter, M. B. (2008). *Speaking in ellipses: The effect of a compensatory style of speech on functional communication in chronic agrammatism*. PhD thesis, Radboud University Nijmegen, Nijmegen, Netherlands.

Russo, N., Larson, C., and Kraus, N. (2008). Audio-vocal system regulation in children with autism spectrum disorders. *Experimental Brain Research*, Volume 188(1):111–124.

Saffran, E., Berndt, R., and Schwartz, M. (1989). The quantitative analysis of agrammatic production: Procedure and data. *Brain and Language*, 37(3):440–479.

Sajwaj, T., Twardosz, S., and Burke, M. (1972). Side effects of extinction procedures in a remedial preschool. *Journal of Applied Behavior Analysis*, 5(2):163–175.

Salis, C. and Edwards, S. (2010). Treatment of written verb and written sentence production in an individual with aphasia: A clinical study. *Aphasiology*, 24(9):1051–1063.

Salt Software LLC (2007). Salt.

Sarno, M. T. (1998). *Acquired aphasia (Third Edition)*. Academic Press, San Diego, CA.

Savitz, D. A. and Olshan, A. (1995). Multiple comparisons and related issues in the interpretation of epidemiologic data. *American Journal of Epidemiology*, 142(9):4.

SaySoft (2007). Annotation.

Schegloff, E. (1982). Discourse as an interactional achievement: Some uses of Ôuh huhÕand other things that come between sentences. *Analyzing discourse: Text and talk*, 71:93.

Schneiderman, B. (1996). The eyes have it: a task by data type taxonomy for informationvisualization. In *IEEE Symposium on Visual Languages*, Boulder, CO.

Schwartz, M., Dell, G., Martin, N., Gahl, S., and Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and Language*, 54(2):228–264.

Scissors, L., Gill, A., Geraghty, K., and Gergle, D. (2009). In cmc we trust: the role of similarity. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 527–536. ACM.

Scissors, L., Gill, A., and Gergle, D. (2008). Linguistic mimicry and trust in text-based cmc. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 277–280. ACM.

Segal, L. B., Oster, H., Cohen, M., Caspi, B., Myers, M., and Brown, D. (1995). Smiling and fussing in seven-month-old preterm and full-term black infants in the still-face situation. *Child Development*, 66(6):1829–1843.

Setlur, V. and Gooch, B. (2004). Is that a smile?: gaze dependent facial expressions. In *Proceedings of the 3rd international symposium on Non-photorealistic animation and rendering*, Annecy, France. ACM.

Shannon, C. and Weaver, W. (1962). *The mathematical theory of communication*, volume 19. University of Illinois Press Urbana.

Sheinkopf, S. J., Mundy, P., Oller, D. K., and Steffens, M. (2000). Vocal atypicalities of preverbal autistic children. *Journal of Autism and Developmental Disorders*, 30(4):345–354.

Shewan, C. and Kertesz, A. (1980). Reliability and validity characteristics of the western aphasia battery (wab). *Journal of Speech and Hearing Disorders*, 45(3):308.

Shill, M. (1979). Motivational factors in aphasia therapy: Research suggestions. *Journal of Communication Disorders*, 12(6):503–517.

S'hiri, S. (2002). Speak arabic please!: Tunisian arabic speakers' linguistic accommodation to middle easterners. In Rouchdy, A., editor, *Language contact and language conflict in Arabic: Variations on a sociolinguistic theme*, pages 149–174. Routledge.

Shuster, L. I. and Ruscello, D. M. (1992). Evoking [r] using visual feedback. *American Journal of Speech-Language Pathology*, 1(May):29–34.

Skinner, E., Wellborn, J., and Connell, J. (1990). What it takes to do well in school and whether i've got it: The role of perceived control in children's engagement and school achievement. *Journal of Educational Psychology*, 82(1):22–32.

Software, Z. (2010). Jbuddy messenger.

Stotland, E. (1969). Exploratory investigations of empathy. *Advances in experimental social psychology*, 4:271–314.

Strand, E., Stoeckel, R., and Baas, B. (2006). Treatment of severe childhood apraxia of speech: A treatment efficacy study. *Journal of Medical Speech Language Pathology*, 14(4):297.

Strik, H. and Cucchiarini, C. (1999). Modeling pronunciation variation for asr: A survey of the literature. *Speech Communication*, 29(2-4):225–246.

Studiocode Business Group (2007). Studiocode business group - supplier of studiocode and stream video analysis and distribution software.

Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press, New York, NY.

Tager-Flusberg, H., Rogers, S., Cooper, J., Landa, R., Lord, C., Paul, R., Rice, M., Stoel-Gammon, C., Wetherby, A., and Yoder, P. (2009). Defining spoken language benchmarks and selecting measures of expressive language development for young children with autism spectrum disorders. *Journal of Speech, Language, and Hearing Research*, 52(3):643.

Takagi, H., Asakawa, C., Fukuda, K., and Maeda, J. (2004). Accessibility designer: visualizing usability for the blind. In *ACM SIGACCESS Accessibility and Computing*, number 77-78, pages 177–184. ACM.

Tartaro, A. (2006). Storytelling with a virtual peer as an intervention for children with autism. Number 84, pages 42–44. ACM.

Tartaro, A. and Cassell, J. (2008). Playing with virtual peers: Bootstrapping contingent discourse in children with autism. In *Proceedings of International Conference of the Learning Sciences*, Utrecht, Netherlands. ACM Press.

Tee, K., Moffatt, K., Findlater, L., MacGregor, E., McGrenere, J., Purves, B., and Fels, S. (2005). A visual recipe book for persons with language impairments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 501–510. ACM.

Tholen, N., Weidner, R., Grande, M., Amunts, K., and Heim, S. (2011). Eliciting dyslexic symptoms in proficient readers by simulating deficits in grapheme-to-phoneme conversion and visuo-magnocellular processing. *Dyslexia*, 17(3):268–281.

Thompson, C., Ballard, K., Tait, M., Weintraub, S., and Mesulam, M. (1997). Patterns of language decline in non-fluent primary progressive aphasia. *Aphasiology*, 11(4-5):297–321.

Toma, C. (2010). Perceptions of trustworthiness online: The role of visual and textual information. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 13–22. ACM.

Tracy, K. and Haspel, K. (2004). Language and Social Interaction: Its Institutional Identity, Intellectual Landscape, and Discipline-Shifting Agenda. *Journal of communication*, 54(4):788–816.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59(236):433–460.

Valstar, M. F., Gunes, H., and Pantic, M. (2007). How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the international conference on Multimodal interfaces*, Nagoya, Aichi, Japan. ACM.

Valstar, M. F., Pantic, M., Ambadar, Z., and Cohn, J. F. (2006). Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Proceedings of the international conference on Multimodal interfaces*, Banff, Alberta, Canada. ACM.

van de Sandt-Koenderman, M., Wiegers, J., and Hardy, P. (2005). A computerised communication aid for people with aphasia. *Disability & Rehabilitation*, 27(9):529–533.

Velleman, S. (2002). Phonotactic therapy. In *Seminars in Speech and Language*, volume 23, pages 35–46.

Viegas, F. B. and Donath, J. S. (1999). Chat circles. In *Proceedings of SIGCHI*, Pittsburgh, Pennsylvania, United States. ACM.

Vredenburg, K., Mao, J., Smith, P., and Carey, T. (2002). A survey of user-centered design practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*, pages 471–478. ACM.

Wang, H. and Fussell, S. (2010). Groups in groups: Conversational similarity in online multicultural multiparty brainstorming. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 351–360. ACM.

Wang, J., Yin, L., and Moore, J. (2007). Using geometric properties of topographic manifold to detect and track eyes for human-computer interaction. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(4):1–20.

Wechsler, D. (1955). Manual for the Wechsler Adult Intelligence Scale.

Weizenbaum, J. (1967). Contextual understanding by computers. *Communications of the ACM*, 10(8):480.

Wetherby, A. M., Prizant, B. M., and Hutchinson, T. A. (1998). Communicative, social/affective, and symbolic profiles of young children with autism and pervasive developmental disorders. *American Journal of Speech-Language Pathology*, 7(May):79–91.

Whyte, W. (1980). *The social life of small urban spaces*. Conservation Foundation, Washington, DC.

Wilson, J., Straus, S., and McEvily, B. (2006). All in due time: The development of trust in computer-mediated and face-to-face teams. *Organizational Behavior and Human Decision Processes*, 99(1):16–33.

Woods, J. J. and Wetherby, A. M. (2003). Early identification of and intervention for infants and toddlers who are at risk for autism spectrum disorder. *Language, Speech, and Hearing Services in Schools*, 34(July 2003):180–193.

Xiao, Z., Poursoltanmohammadi, A., and Sorell, M. (2008). Video motion detection beyond reasonable doubt. In *Proceedings of the international conference on Forensic applications and techniques in telecommunications, information, and multimedia and workshop*, Adelaide, Australia. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Yngve, V. (1970). On getting a word in edgewise. In *Chicago linguistic society - 70*, pages 567–578.

Yoder, P., Camarata, S., and Gardner, E. (2005). Treatment effects on speech intelligibility and length of utterance in children with specific language and intelligibility impairments. *Journal of Early Intervention*, 28(1):34.

Zacks, J., Tversky, B., and Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *J Exp Psychol Gen*, 130(1):29–58.

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2007). A survey of affect recognition methods: audio, visual and spontaneous expressions. In *Proceedings of the international conference on Multimodal interfaces*, Nagoya, Aichi, Japan. ACM.

Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458.